Data Analysis III

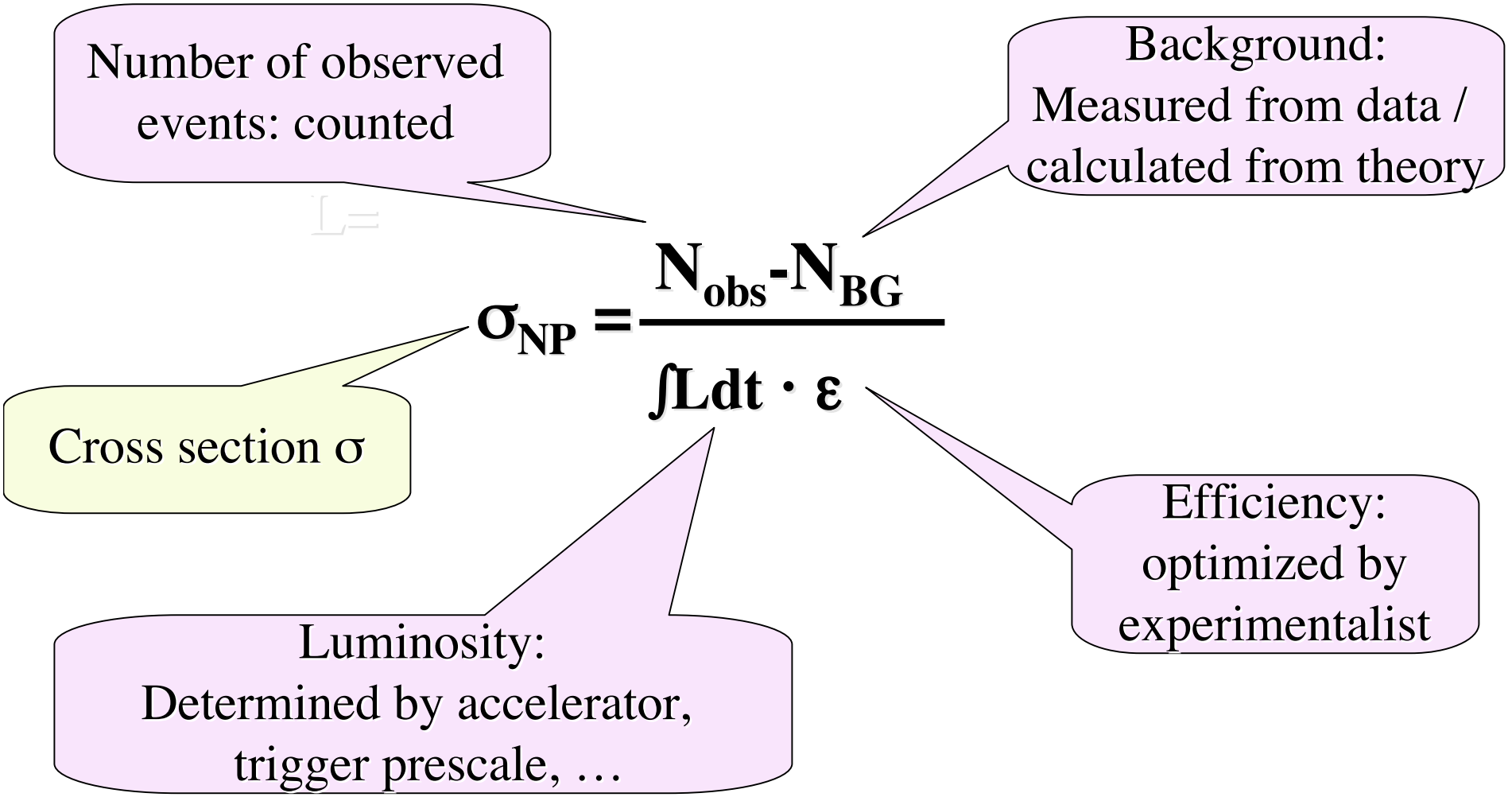
Beate Heinemann

UC Berkeley and Lawrence Berkeley National Laboratory

Outline

- Lecture I:
  - Measuring a cross section
    - focus on acceptance
- Lecture II:
  - Measuring a property of a known particle
- Lecture III:
  - Searching for a new particle
    - focus on backgrounds

Search for New Particles: Experimentally



- Exactly like with measuring the cross section…

But we need to observe first!

- When we don't know if a particle exists our first question is: "Does it exist?"
- => significance of signal
  - I.e. how consistent is the number of observed events with the number of background events?
  - Background expectation: N_{BG}
    - Expect it to fluctuate statistically by $\delta N_{BG} \sim \sqrt{N_{BG}}$
  - Signal expectation: N_{Signal}
  - Statistical Significance: $N_{Signal}/\delta N_{BG} \sim N_{Signal} / \sqrt{N_{BG}}$
    - Often called $S/\sqrt{B}$

in Gaussian limit

	evidence	observation
significance	**3σ**	**5σ**
Probability of stat. fluctuation	**0.3%**	**5.7×10^{-8}**

In real life: systematic uncertainties also contribute to δN_{BG}

Search analyses

- Primary focus is background estimate
  - Determines whether or not an observation can be made
  - Cuts for background reduction studied often using benchmark New Physics scenario
    - Also model-independent analyses..
- Secondary focus is acceptance/efficiency determination: required only
  - when putting an upper limit on a cross section
  - when measuring the cross section of the observed new particle
    - Need to know what it is though
    - Or quote cross section for some effective cuts

Example Analyses

- SUSY:
  - Squarks/gluinos → jets + E_T (+leptons)
- Higgs:
  - Higgs -> WW
- High Mass Resonance
  - Z' (if I have time)

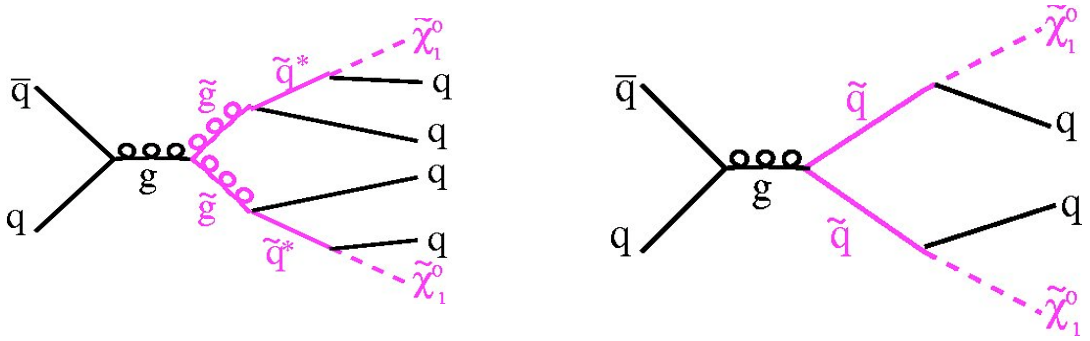
Backgrounds

- Ideally you get the backgrounds to be small
  - The smaller they are the less well you need to know them
- Estimates based on
  - Data only
    - E.g. lepton fake rates
  - Monte Carlo only
    - For well known electroweak processes
  - Monte Carlo / Data hybrid
    - For e.g. W/Z+jets or W/Z+b-jets

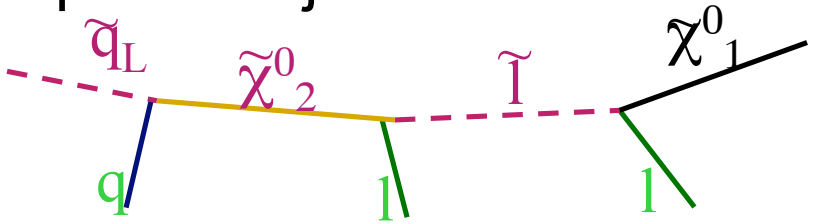
Squarks/Gluinos → Jets + MEt (+ leptons)

SUSY at the LHC

- Cross section **much** higher than at Tevatron, e.g.
  - for $m(\tilde{g})$=400 GeV: $\sigma_{LHC}(\tilde{g}\tilde{g})/\ \sigma_{Tevatron}(\tilde{g}\tilde{g})\approx 20{,}000$
  - for $m(\tilde{q})$=400 GeV: $\sigma_{LHC}(\tilde{g}\tilde{g})/\ \sigma_{Tevatron}(\tilde{g}\tilde{g})\approx 1{,}000$
    - Since there are a lot more gluons at the LHC (lower x)

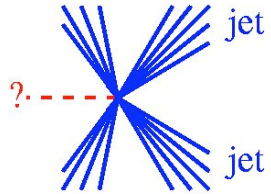


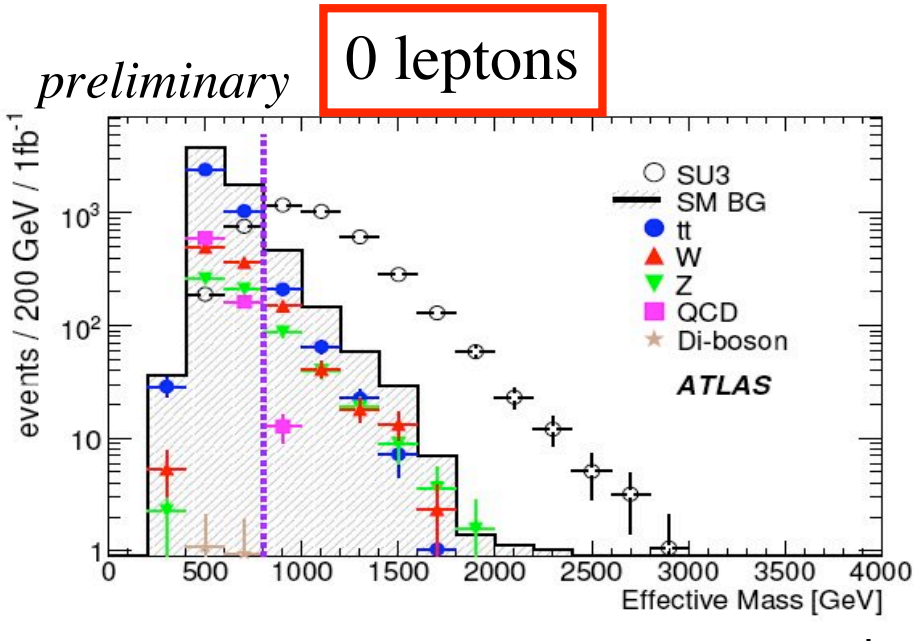
- At higher masses more phase space to decay in cascades
  - Results in additional leptons or jets

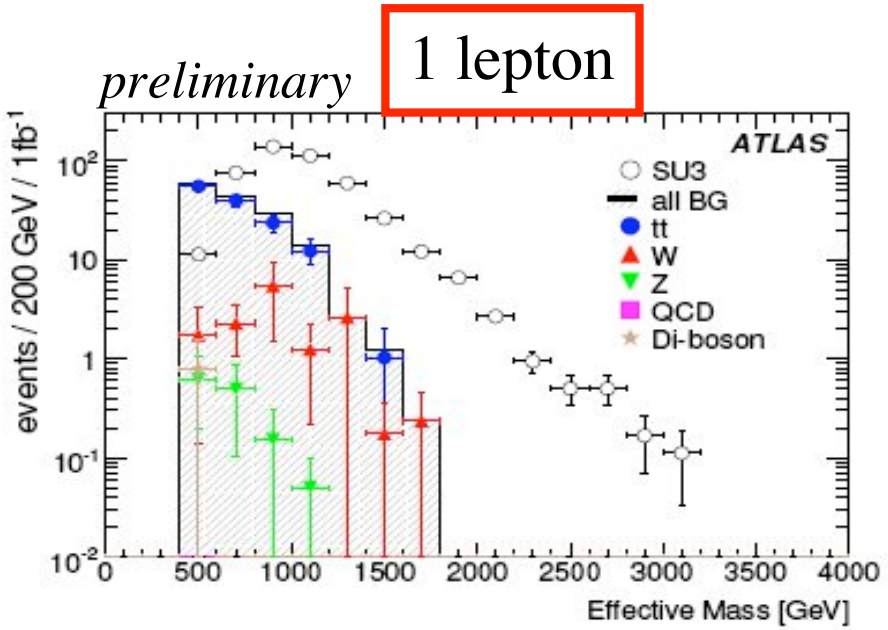


SUSY at the LHC

- Example: m($\tilde{q}$)~600 GeV, m($\tilde{g}$)~700 GeV
- Require 4 jets, large missing E_T and 0 or 1 lepton





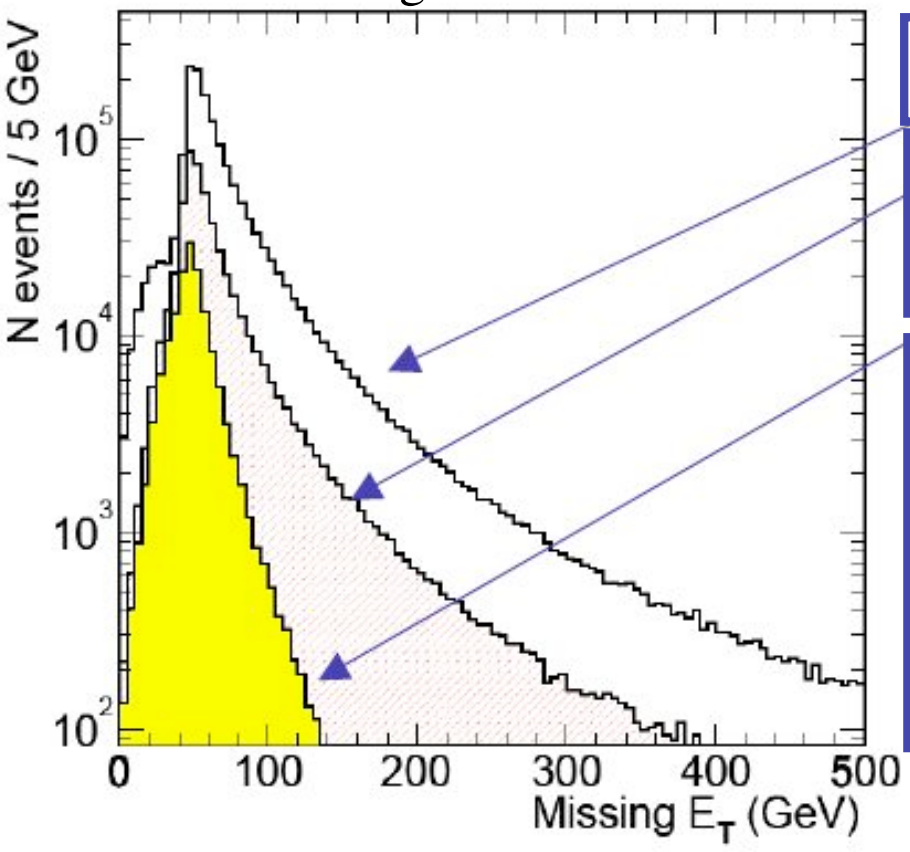


- Similar and great (!) sensitivity in both modes
- Main backgrounds: top, W/Z+jets, QCD multi-jet

But how do we know the backgrounds!?!

Instrumental Backgrounds

from Avi Yagil



After selection of "good runs"

After requirement of vertex in tracker

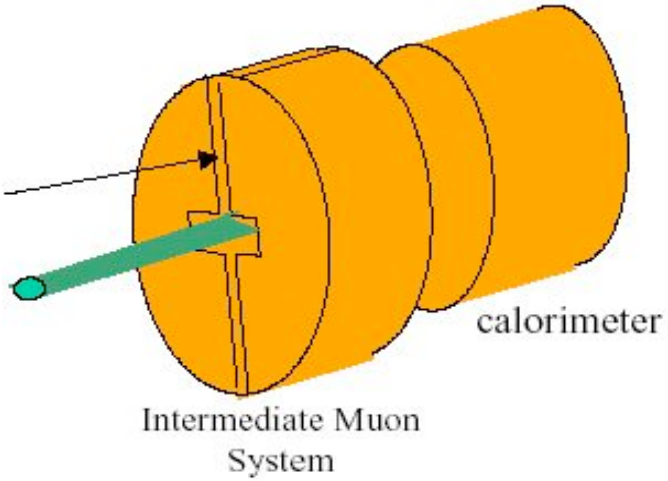
After "clean-up" cuts:
- event EM fraction > 0.1
- event charge fraction > 0.1
- ≥1 jet with E_T>10 GeV
- $E_{total} < \sqrt{s}$

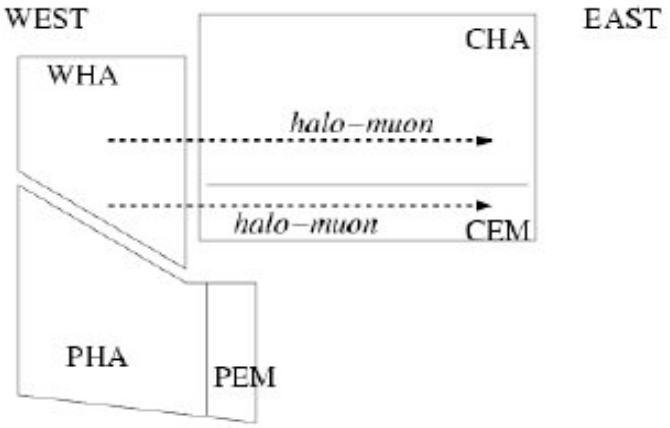
- Missing ET distribution subject to many experimental effects
  - If anything goes wrong it will affect missing ET

Sources of Instrumental Background

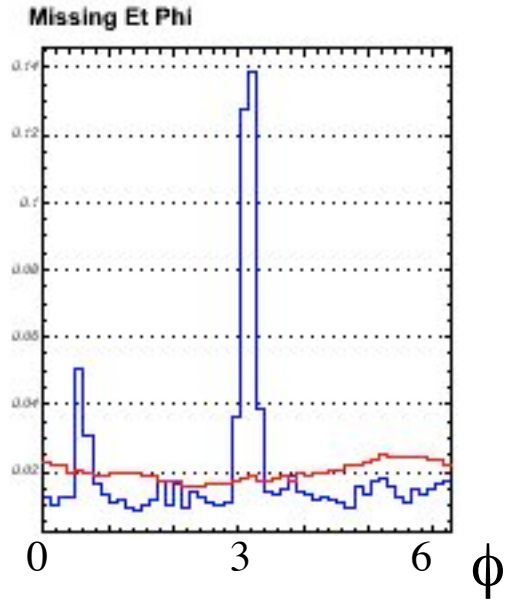
- Calorimeter Noise
  - Hot cells / coherent noise
    - Usually localized and can be rejected
- Calorimeter dead regions
  - Should only happen rarely in some runs
    - Should be removed by DQ criteria
- Cosmic rays and beam halo muons showering hard in calorimeter
  - Usually have no vertex but can overlap with MinBias event
    - Then have small tracking activity compared to calorimeter activity
  - Shower often only in hadronic calorimeter
- Example handles:
  - Track/calorimeter matches
  - Is direction of missing energy uniform?

Beam-Halo Muon Background

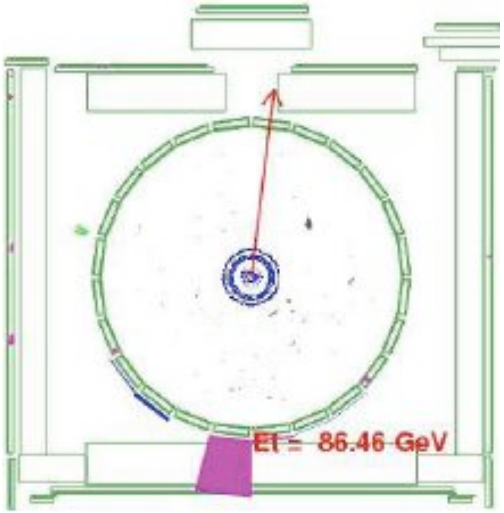


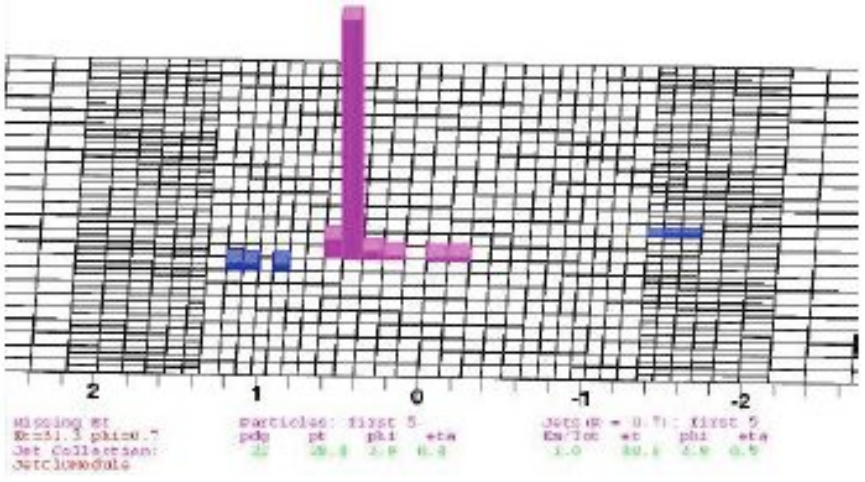


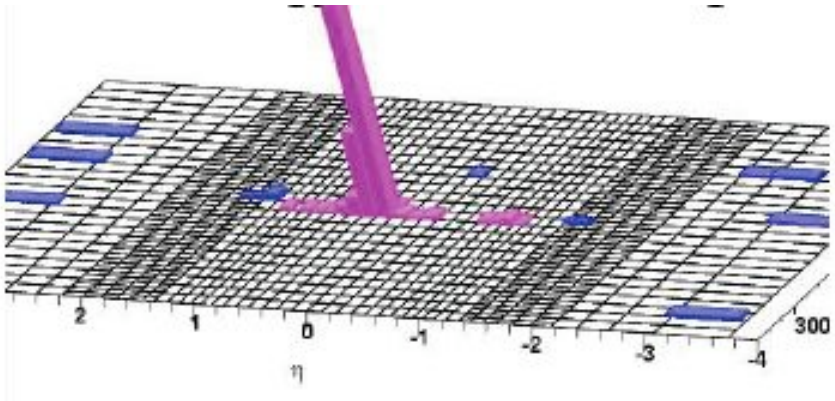
- Muon that comes from beam and goes through shielding
- Can cause showers in calorimeters
  - Shower usually looks not very much like physics jet
    - Often spike at certain azimuthal angles: π
  - But there is lots of those muons!
  - Can even cause problem for trigger rate



Some Cosmics and Beam-halo events







- Bigger problem for mono-jet than for multi-jet searches
- Can use
  - topological filters to reject events
  - Track matching calorimeter cluster

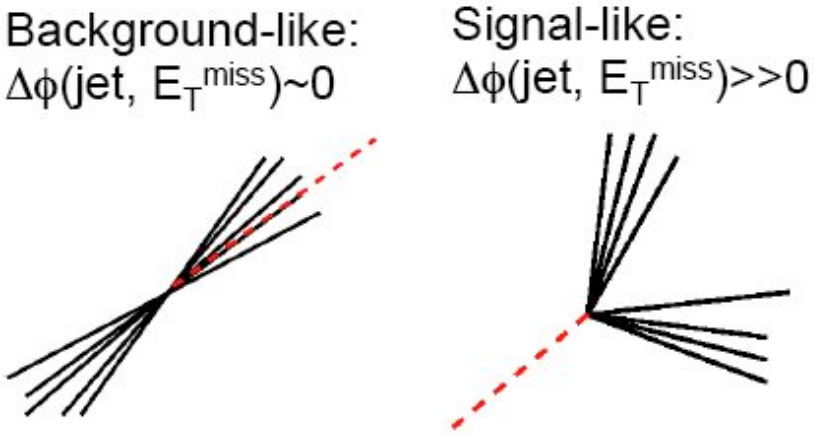
Physics Backgrounds

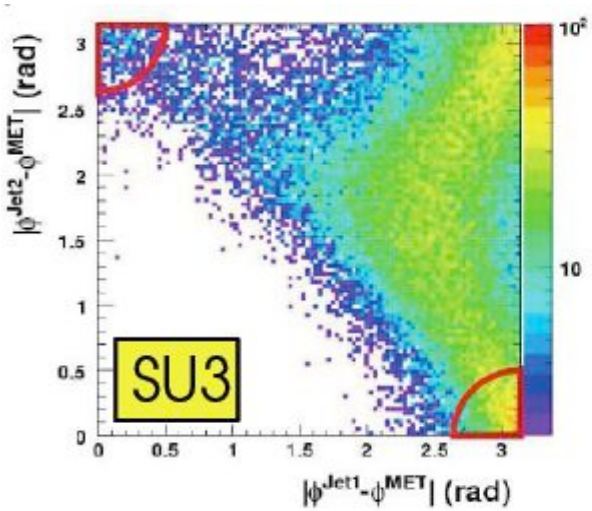
- QCD multi-jet (mosty for 0-lepton case)
  - Missing E_T due to
    - Poor jet resolution / cracks in calorimeters
    - Neutrino momentum in semi-leptonic b/c- decays
- W/Z+jets
  - Missing E_T due to ν's from $Z \rightarrow \nu\nu$, $W \rightarrow l\nu$
- Top
  - Missing E_T due to ν's from $tt \rightarrow WbWb \rightarrow l\nu + X$

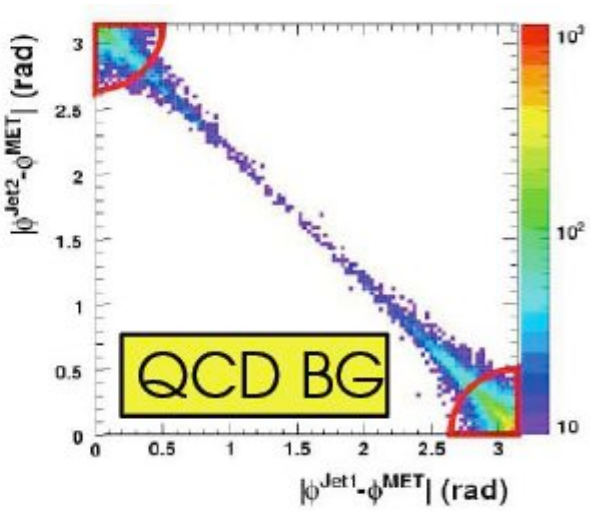
How do we estimate them?

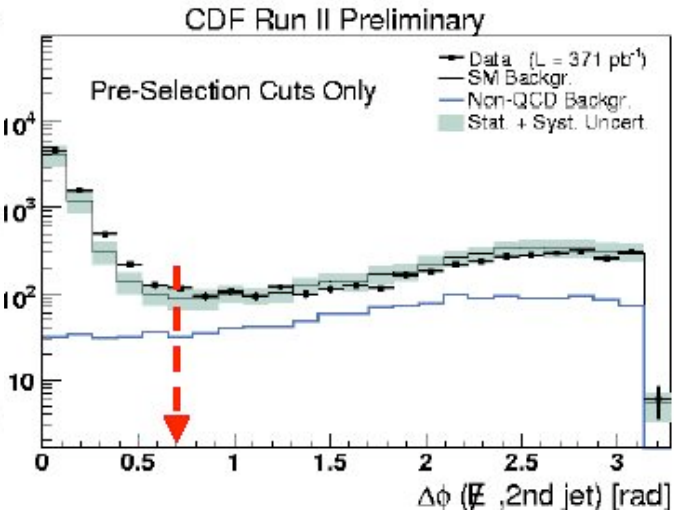
QCD Multi-jet

- Require large $\Delta\phi$
  - Between missing E_T and jets and between jets
  - Suppresses QCD dijet background due to jet mismeasurements



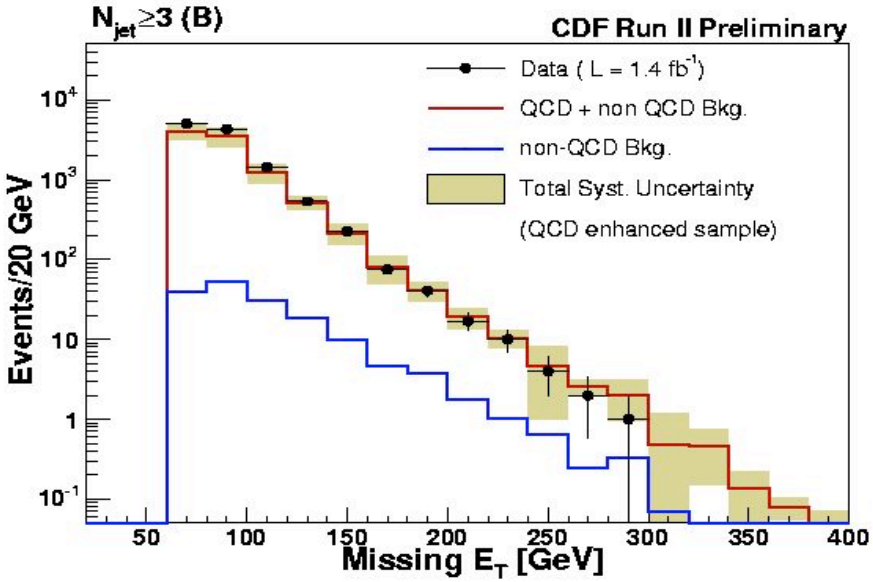


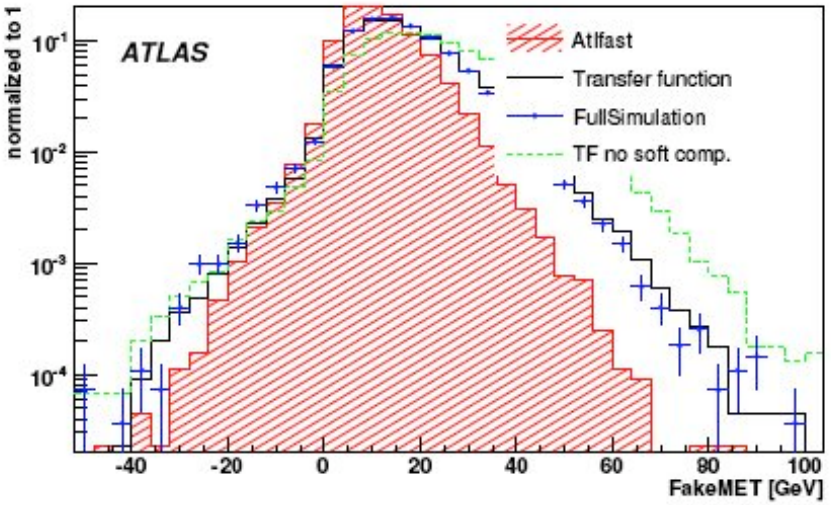




Methods to estimate remaining QCD multi-jet Background

1. CDF uses MC
   - Validate in region of low $\Delta\Phi$and low MET
   - Extrapolate to large using MC
   - Problem:
     - Relies on full MC simulation which can take “forever”
2. Parameterize truth jets with response function from full simulation
   - Validate against full simulation
   - Validate in region of lower MET
   - Advantage:
     - Do not need to simulate as many events
   - Need to make sure though that parameterization is really working

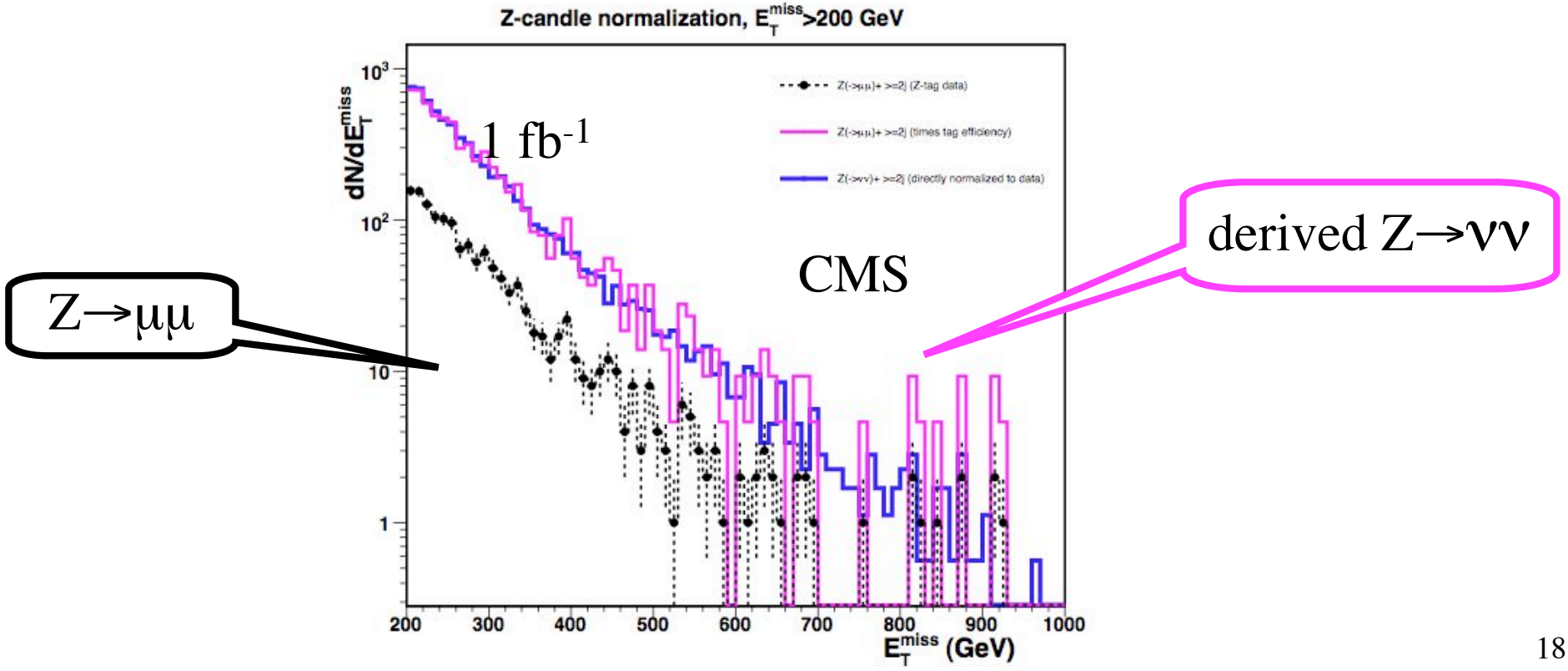




Using Z(→ll)+jets for estimating W/Z+jet background

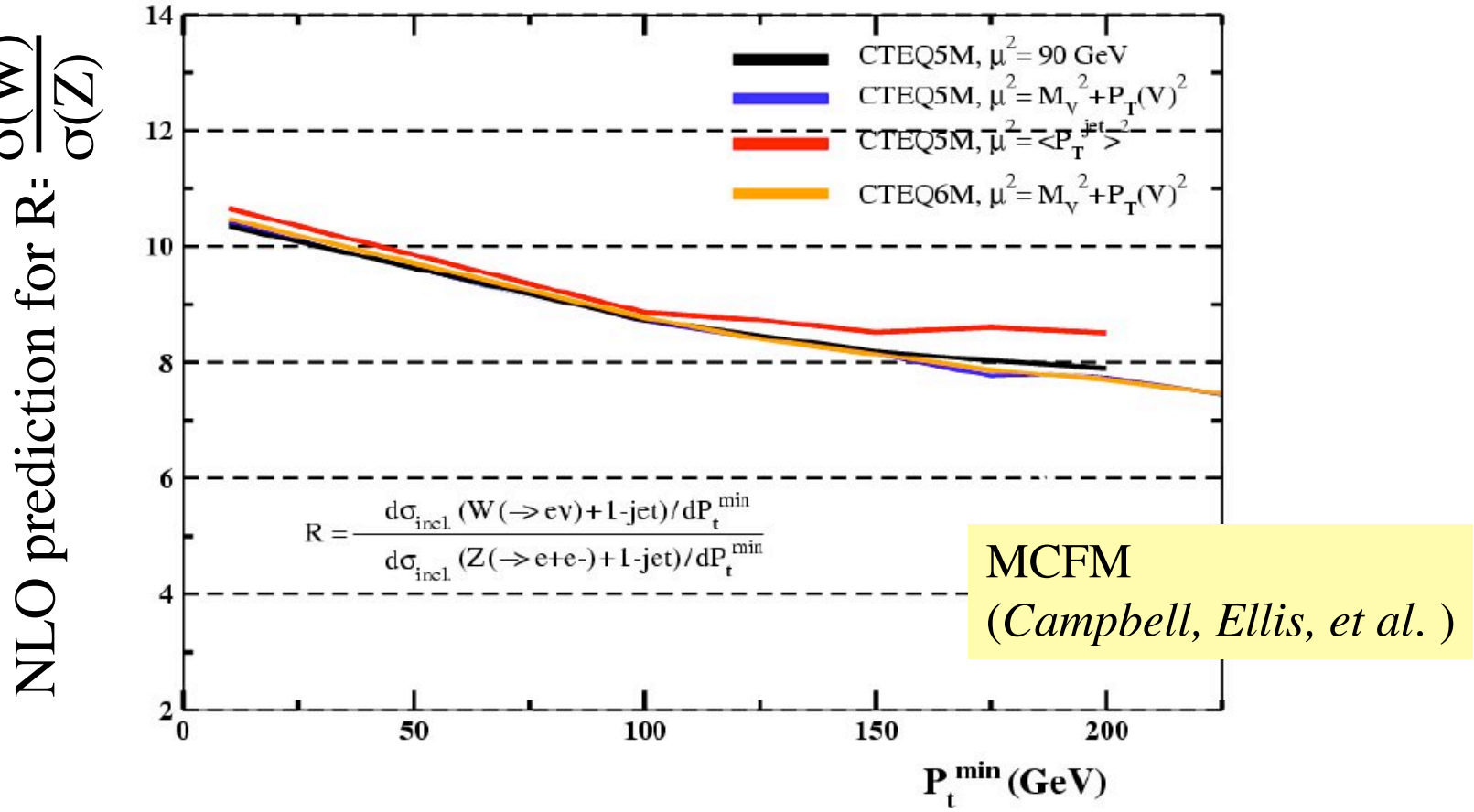
- Use Z(→ll)+jets to extrapolate to Z(→νν)+jets
  - ME_T ~ $p_T(Z)$

$$N_{Z\to\nu\bar{\nu}}(E_T^{miss}) = N_{Z\to\ell^+\ell^-}(p_T(\ell^+\ell^-)) \times c_{Kin}(p_T(Z)) \times c_{Fidu}(p_T(Z)) \times \frac{Br(Z\to\nu\bar{\nu})}{Br(Z\to\ell^+\ell^-)},$$



W+jets background estimate

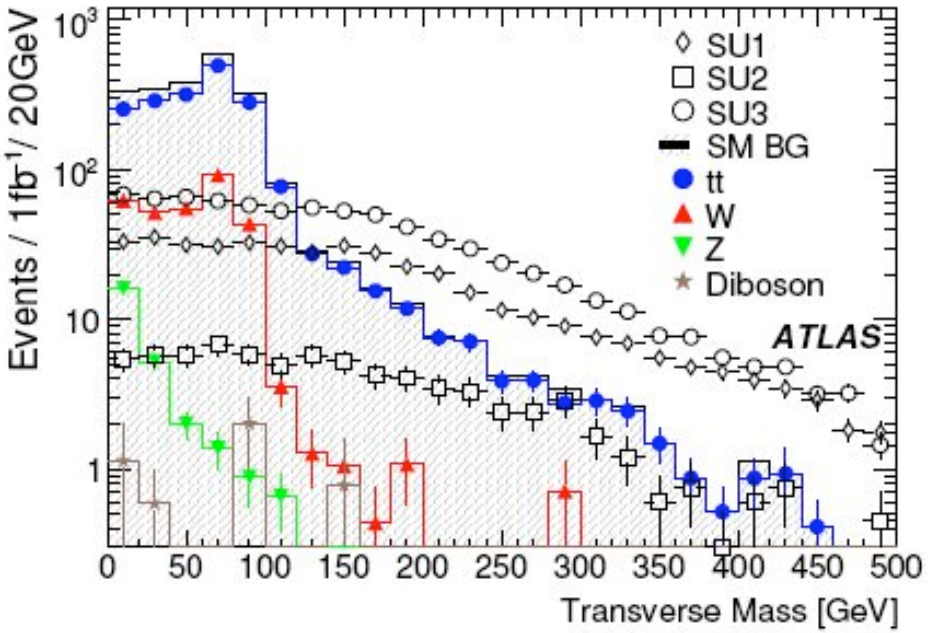
- Use Z->ll +jets also for this background too
  - Rely on theoretical prediction for W+jets vs Z+jets
    - This is well known though (<15%)!

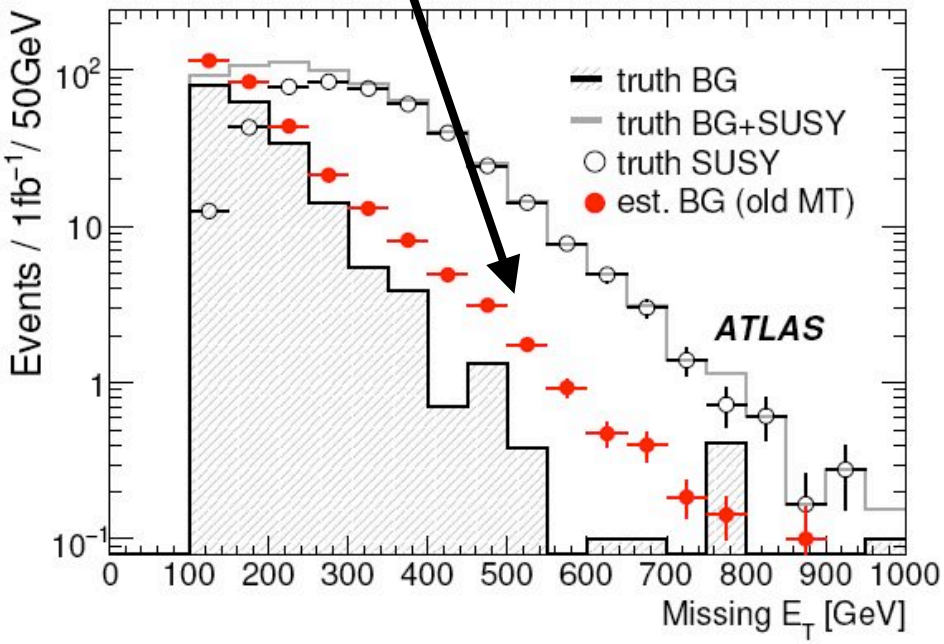


MCFM
(*Campbell, Ellis, et al.*)

Top and W+jets background estimate

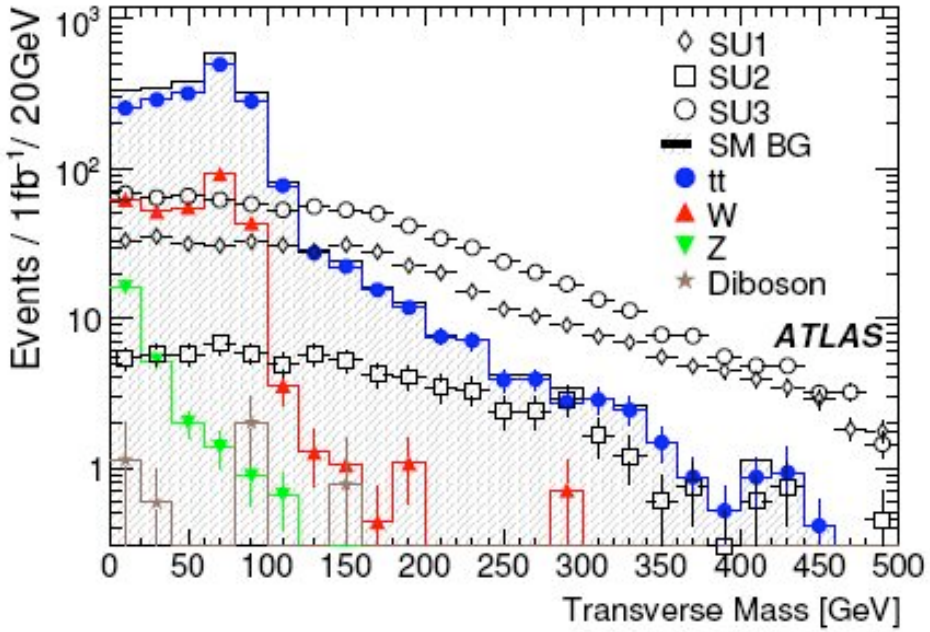
- Use region of low $m_T(W)$
  - Extrapolate to signal region using MC
  - But may be contaminated by SUSY => overestimate BG
    - depending on specifics of model

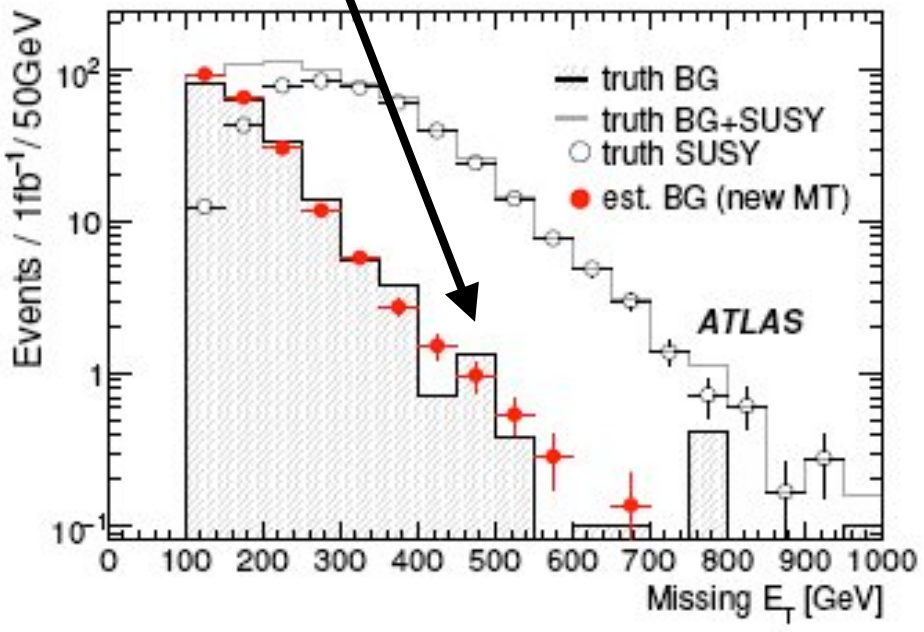




Top and W+jets background estimate

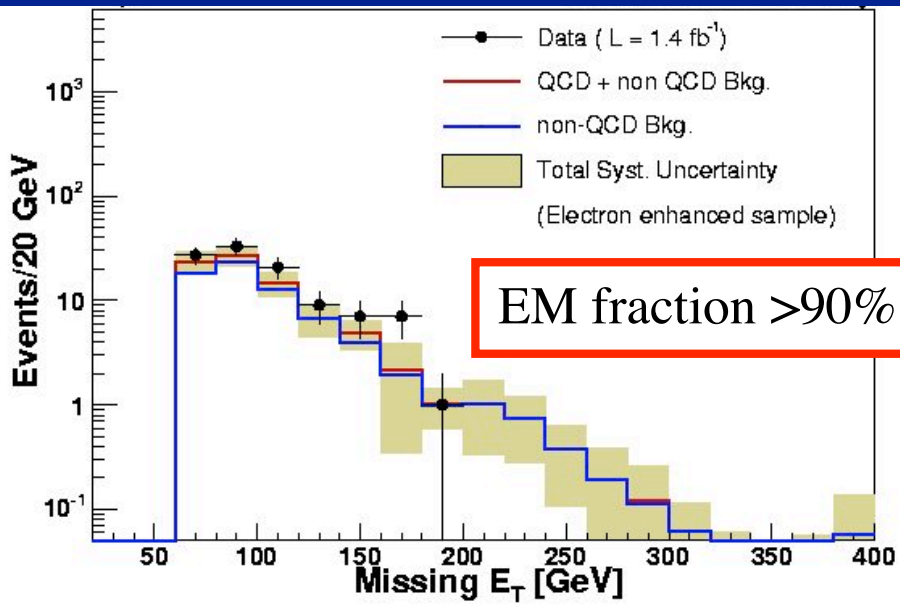
- Use region of low m_T(W)
  - Extrapolate to signal region using MC
  - But may be contaminated by SUSY => overestimate
    - depending on specifics of model
  - Can attempt “SUSY background subtraction” to correct for it

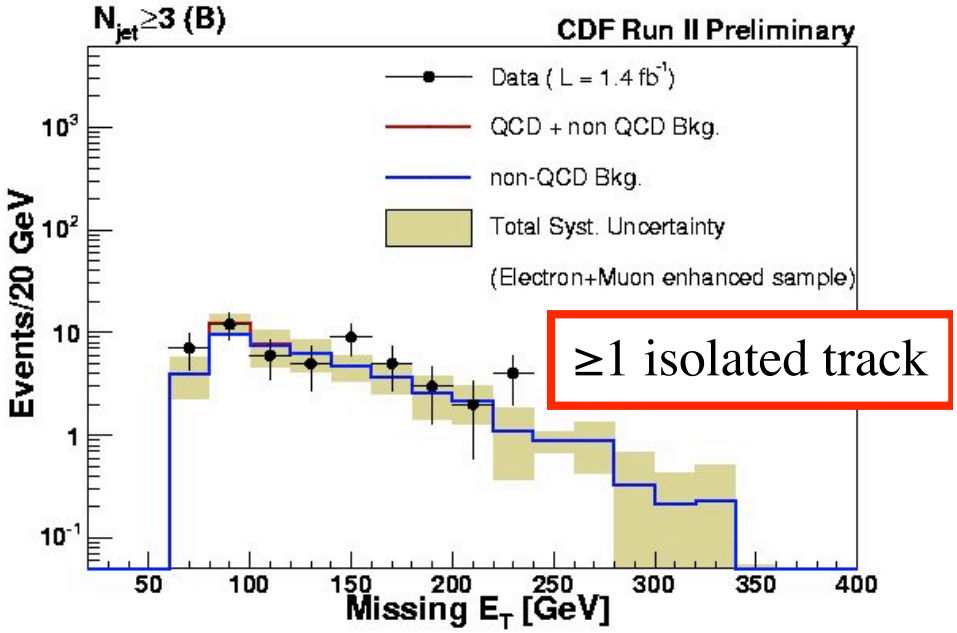




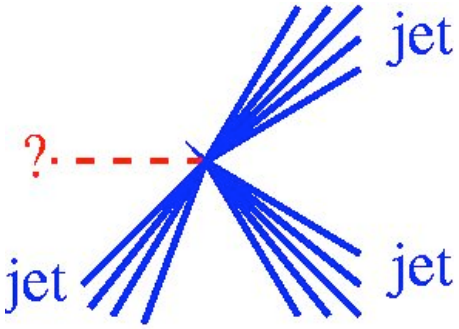
W+jets, Z+jets and Top background

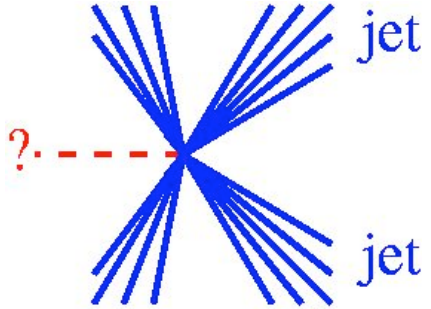
- Checks at Tevatron 0-lepton analysis
  - Background sources:
    - W/Z+jets, top
    - Suppressed by vetoes:
      - Events with jet with EM fraction>90%
        - » Rejects electrons
      - Events with isolated track
        - » Rejects muons, taus and electrons
  - Define control regions:
    - W/Z+jets, top
      - Make all selection cuts but invert lepton vetoes
    - Gives confidence in those background estimates
      - Modeled using Alpgen MC
      - Cross sections determined using NLO calculation
- May not work at LHC due to expectation of large cascade decays

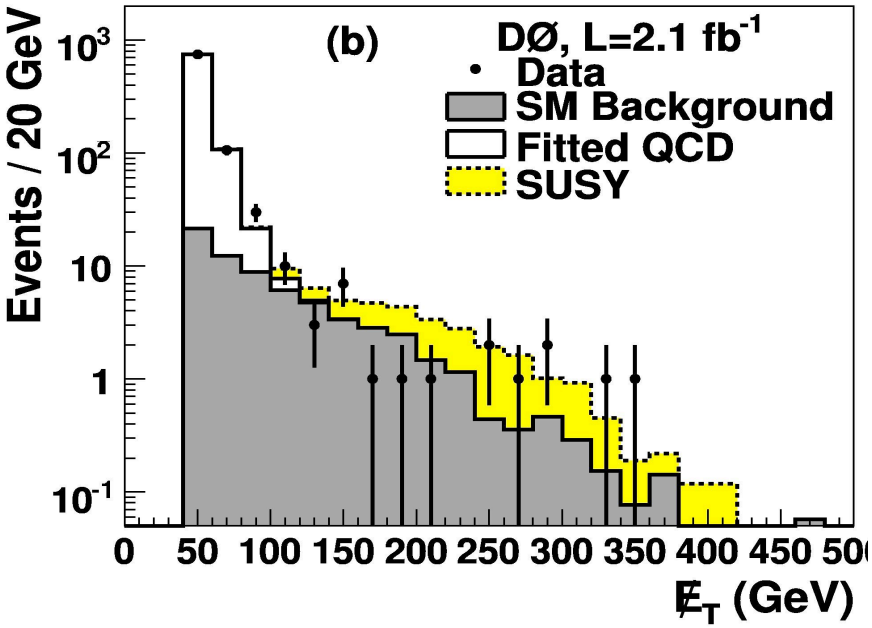


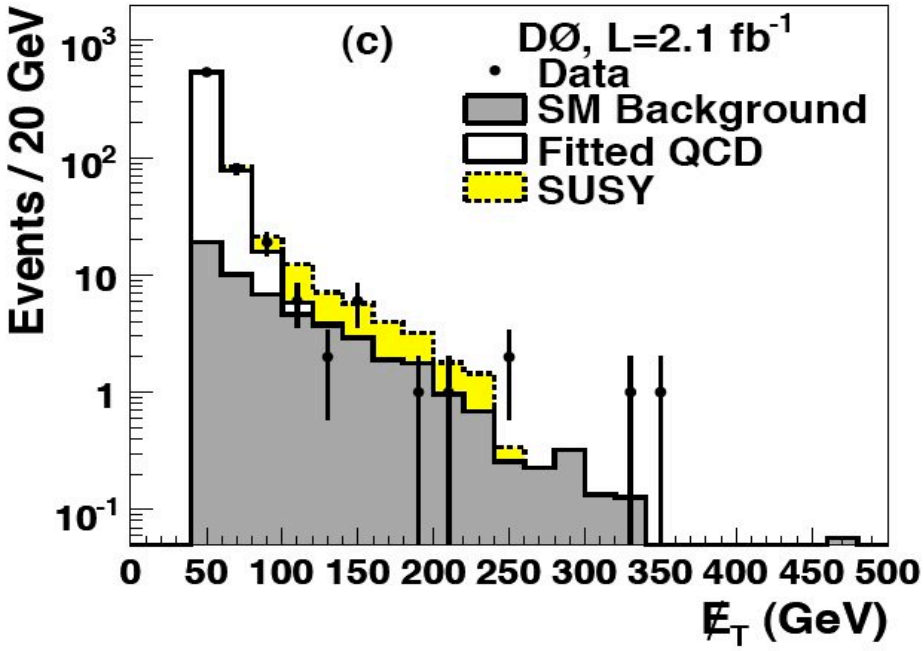


Final Analysis Plots at the Tevatron



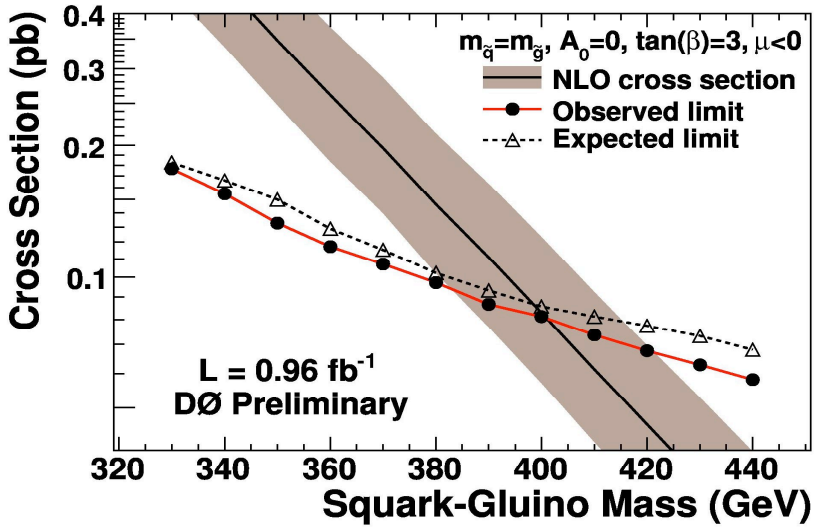


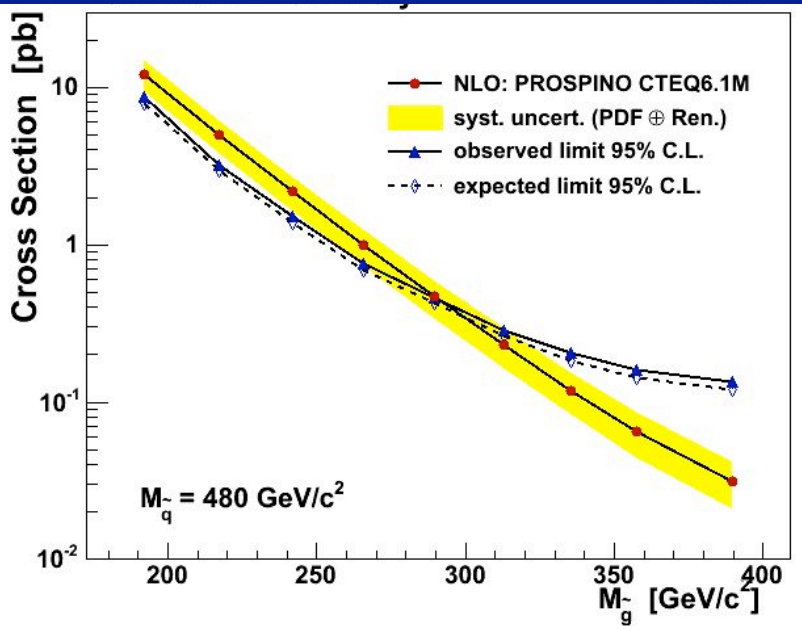




Data agree with background estimate => derive limits

Cross Section Limits

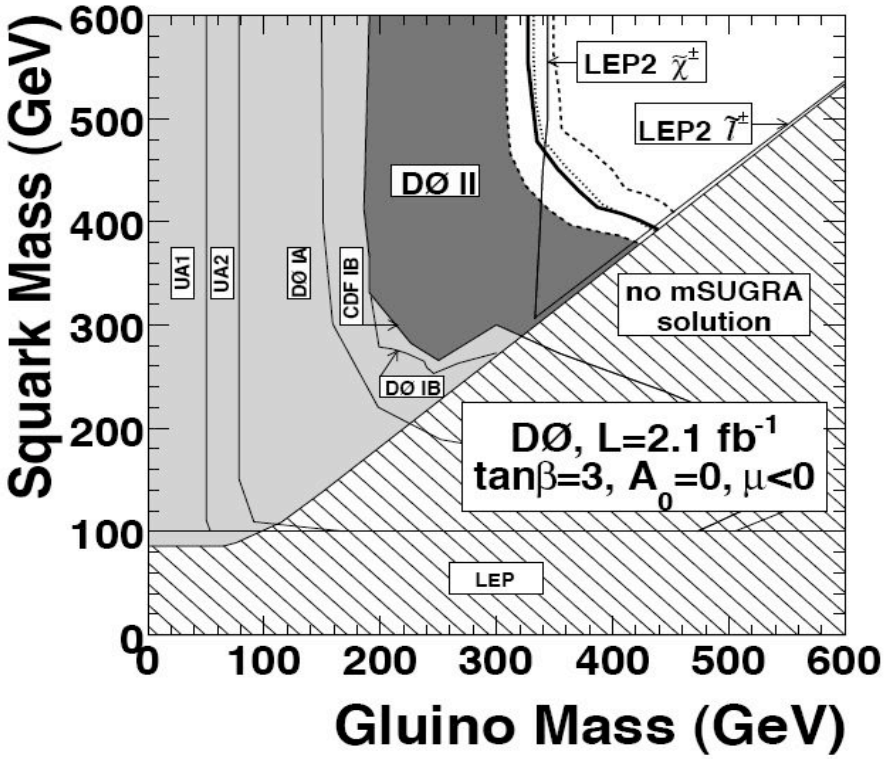


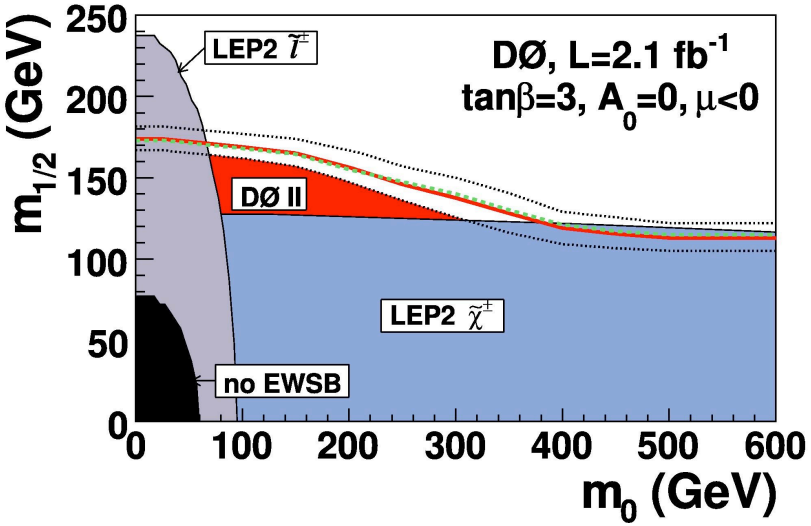


- No excess in data
  - Evaluate upper limit on cross section
  - Find out where it crosses with theory
- Theory has large uncertainty: ~30%
  - Crossing point with theory lower bound ~ represents limit on squark/gluino mass

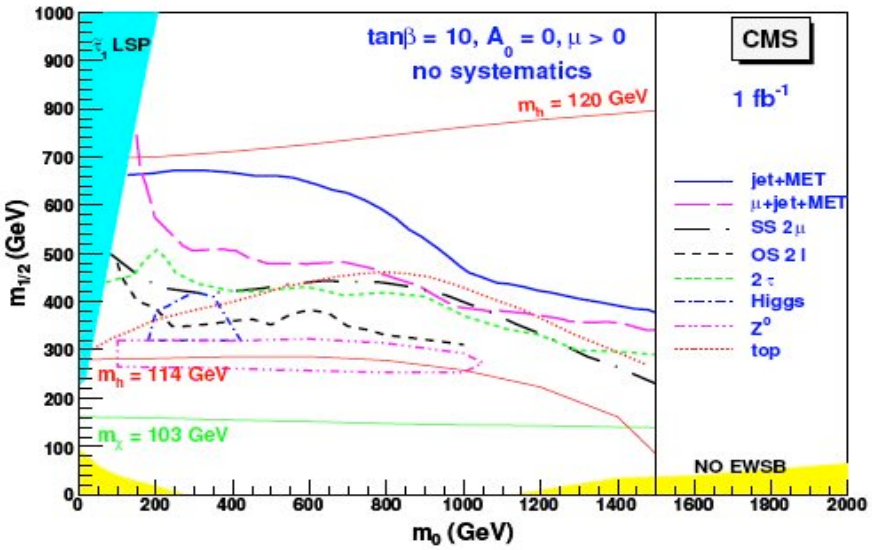
Squark and Gluino Mass Limits

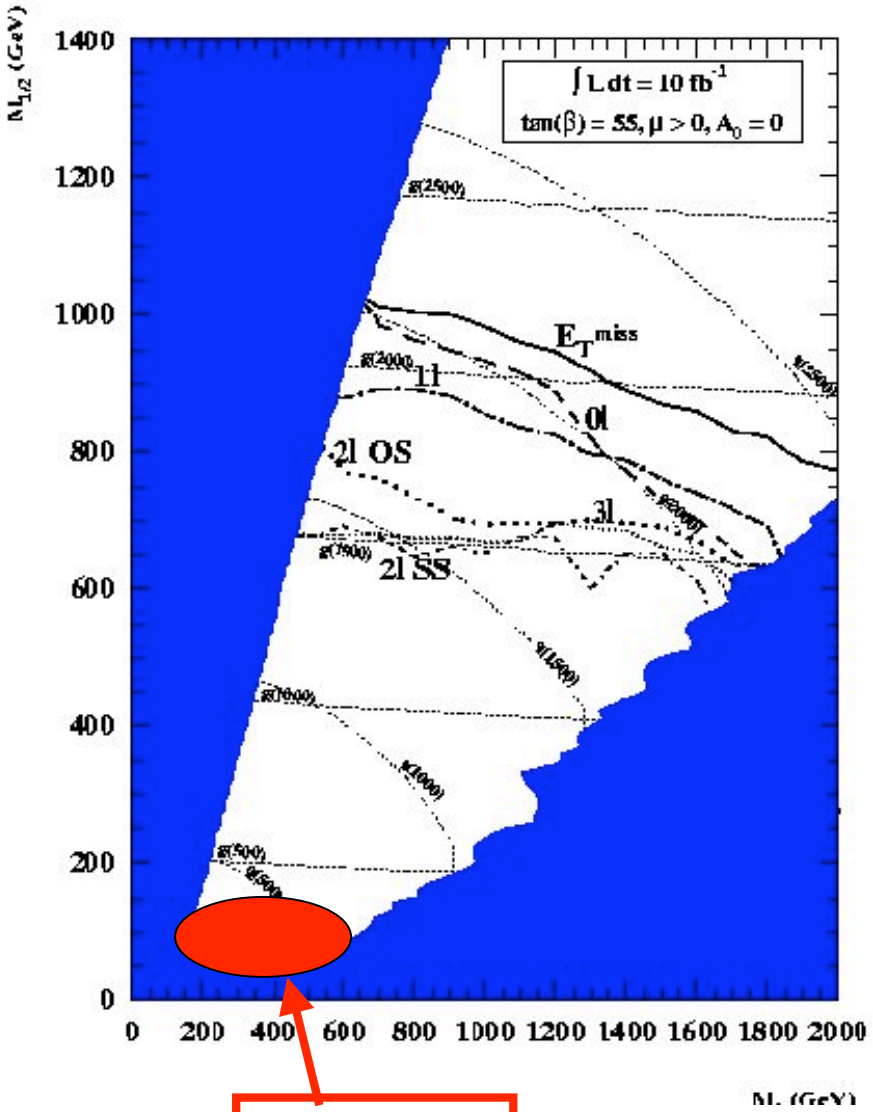
- Set constraints on masses at EWK scale:
  - M($\tilde{g}$)>308 GeV
  - M($\tilde{q}$)>379 GeV
- Can also represented in terms of GUT scale parameters
  - Within constrained models





LHC SUSY Discovery Reach

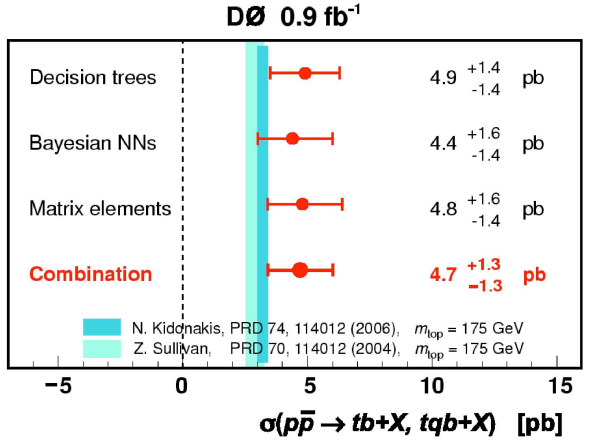




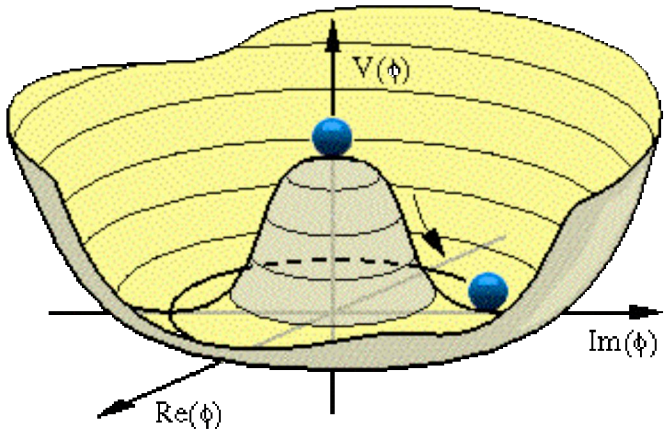
- With 1 fb^{-1}:
  - Sensitive to m($\tilde{g}$)<1000 GeV/c^2
- With 10 fb^{-1}:
  - Sensitive to m($\tilde{g}$)<1800 GeV/c^2
- Amazing potential!
  - If data can be understood
  - If current MC predictions are ≈ok

Some Remarks on Advanced Analysis Techniques

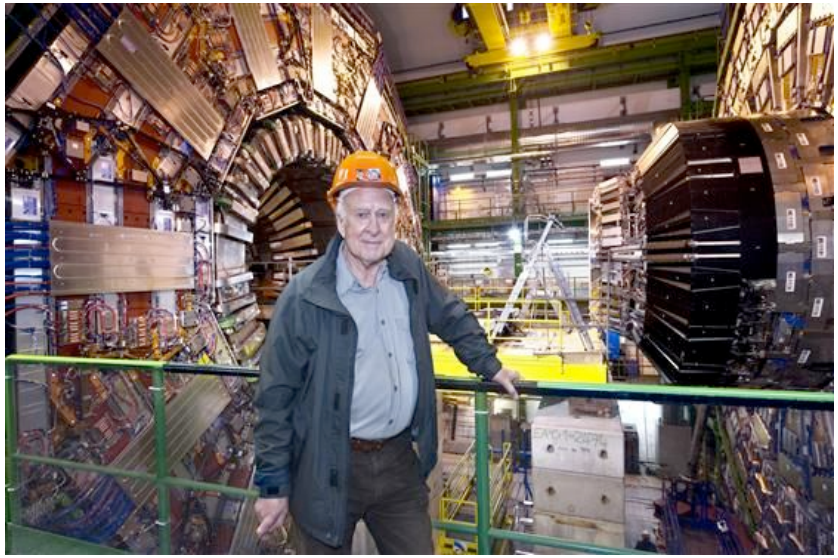
- **Quite a few techniques available:**
  - Neural Network, Likelihood, Boosted Decision Tree, Matrix Element, …
  - No clear winner has yet been identified
    - Generally NN least transparent probably



- **Why do we trust them less than simple analyses?**
  - Simple kinematic quantities can be calculated at NLO by theorists while e.g. NN distribution cannot
    - Gives confidence, good cross-check!
  - Techniques exploit correlations between variables
    - Harder to understand if the MC models correlations correctly
    - More validation needed (=> analysis takes longer)
  - Less transparent
    - Worry is always that it exploits some MC feature that does not reflect the data
- **Can and has been done of course though**
  - But only in mature experiments

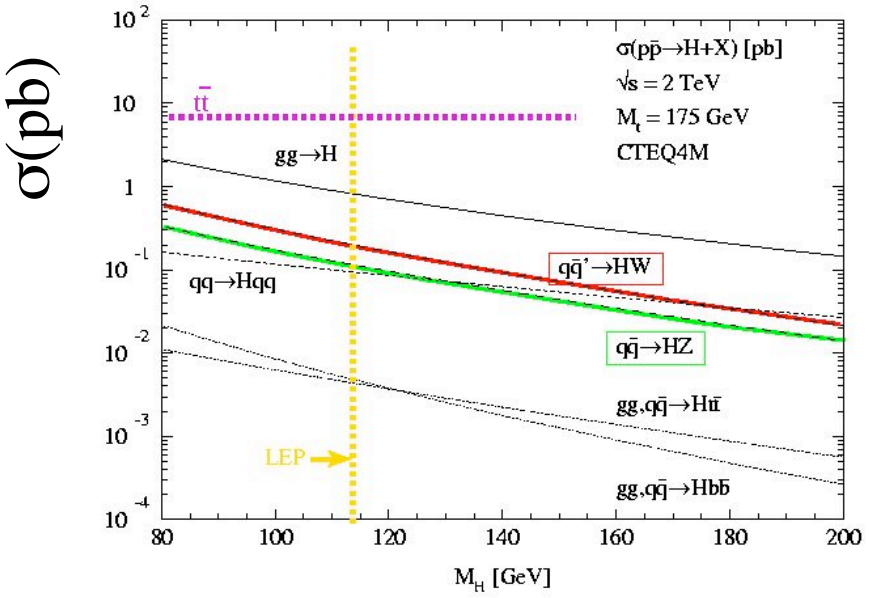


The Higgs Boson

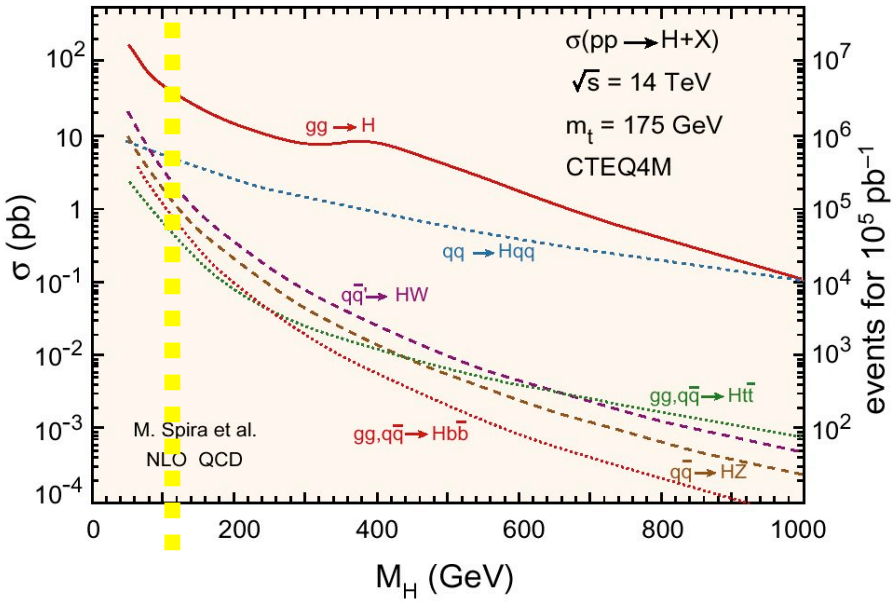


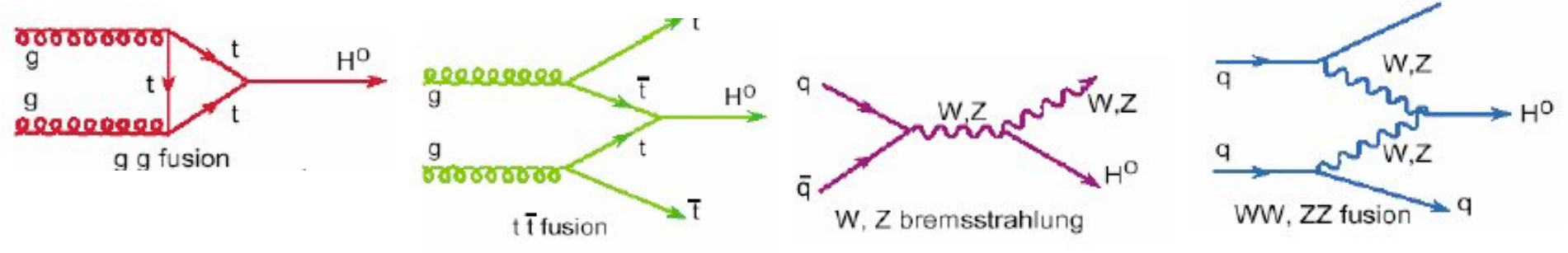
Higgs Production: Tevatron and LHC

Tevatron



LHC

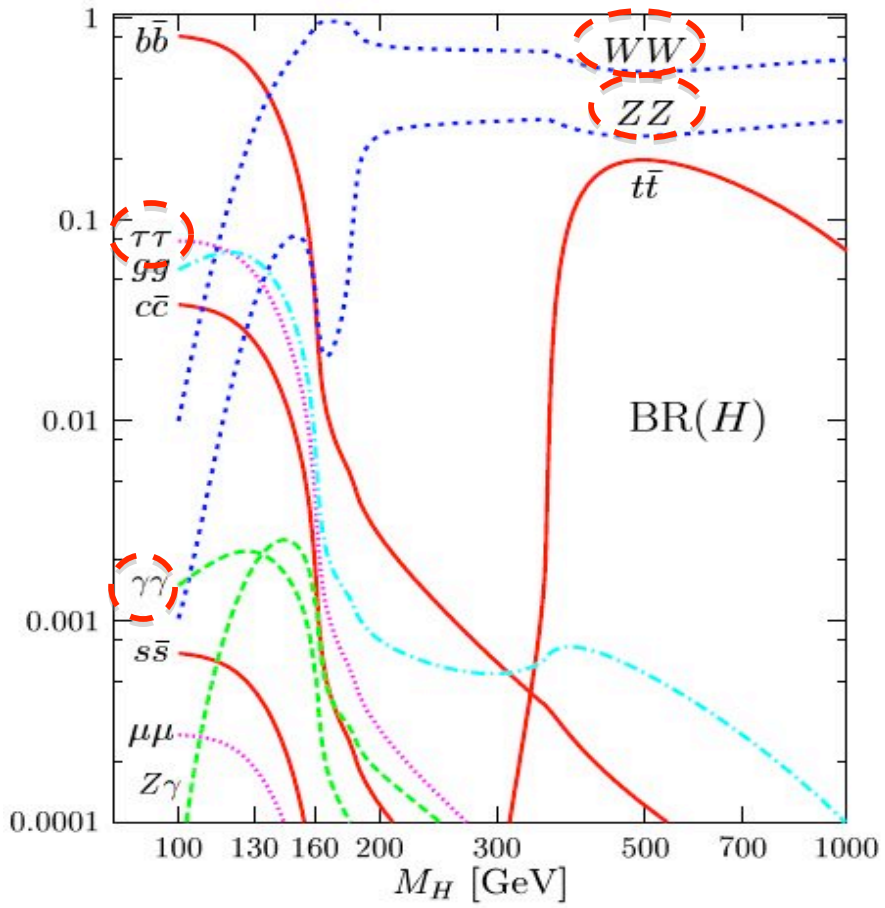




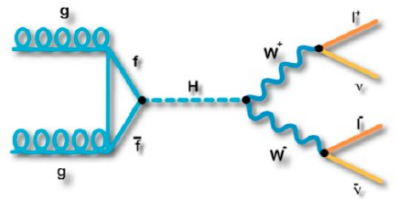
dominant: gg→ H, subdominant: HW, HZ, Hqq

Higgs Boson Decay

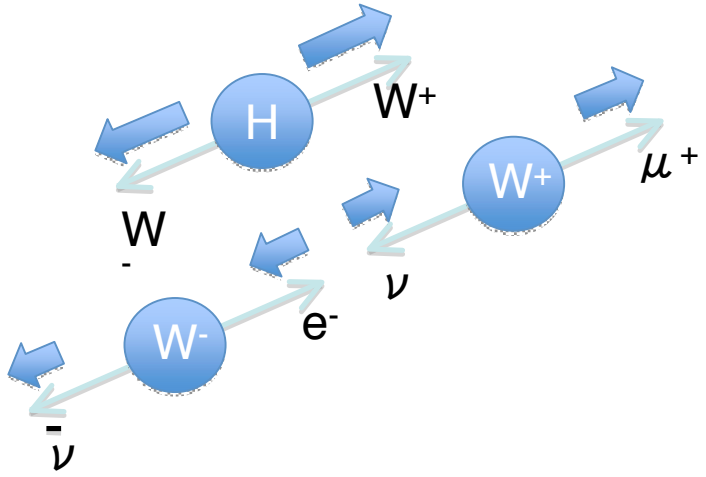
- Depends on Mass
- M_H<130 GeV/c^2:
  - $b\bar{b}$ dominant
  - WW and $\tau\tau$ subdominant
  - $\gamma\gamma$ small but useful
- M_H>130 GeV/c^2:
  - WW dominant
  - ZZ cleanest

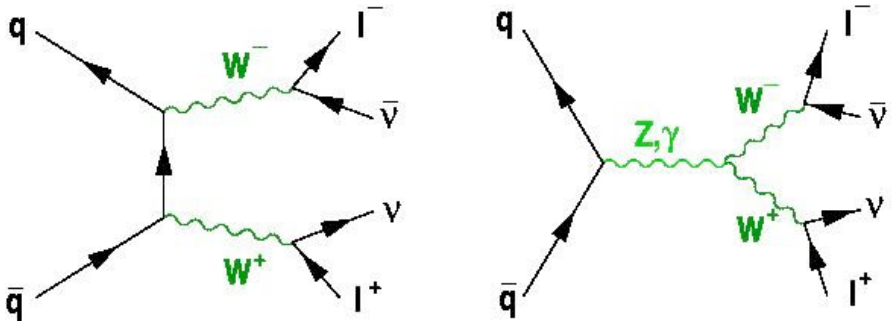


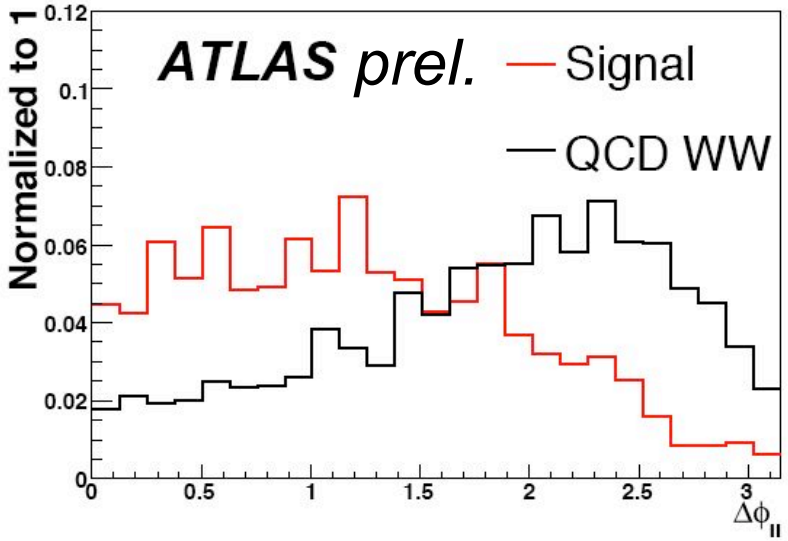
$H \rightarrow WW^{(*)} \rightarrow l^+l^-\bar{\nu}\nu$



- Higgs mass reconstruction impossible due to two neutrinos in final state
- Make use of spin correlations to suppress WW background:
  - Higgs is scalar: spin=0
  - leptons in $H \rightarrow WW^{(*)} \rightarrow l^+l^-\nu\nu$ are collinear
- Main background:
  - WW production

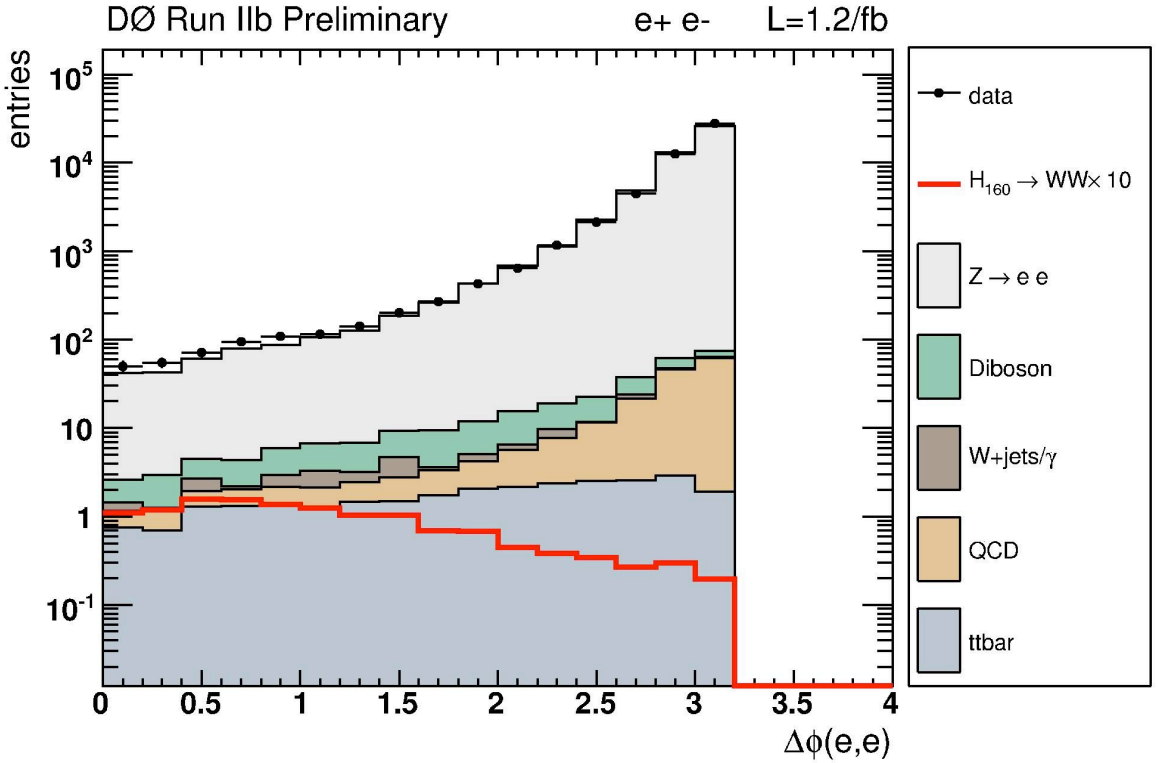






$H \rightarrow WW^{(*)} \rightarrow l^+l^-\nu\nu$ (l=e,μ)

- **Event selection:**
  - 2 isolated e/μ :
    - p_T > 15, 10 GeV
  - Missing E_T >20 GeV
  - Veto on
    - Z resonance
    - Energetic jets
- **Main backgrounds**
  - SM WW production
  - Top
  - Drell-Yan
  - Fake leptons



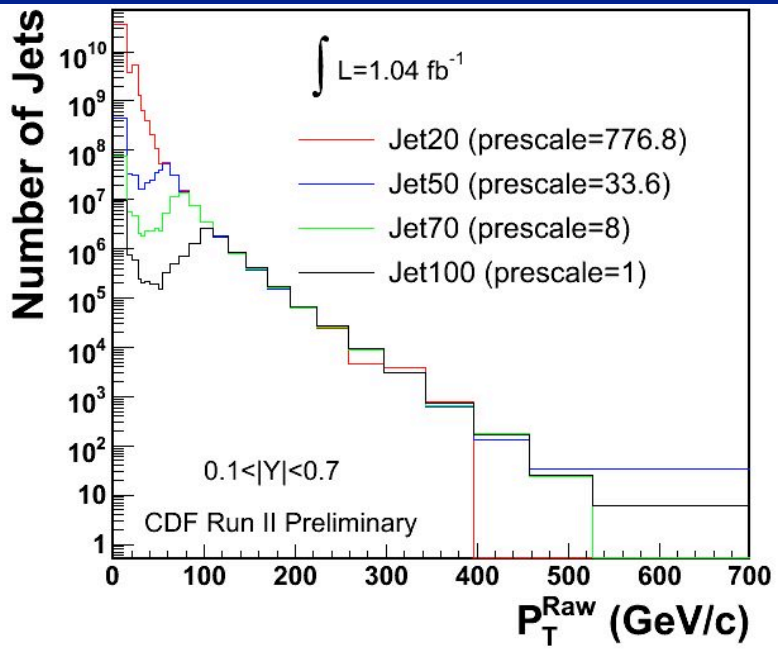
- **Plot everything under the sun**
  - to convince yourself you have the background right

Jets faking Electrons

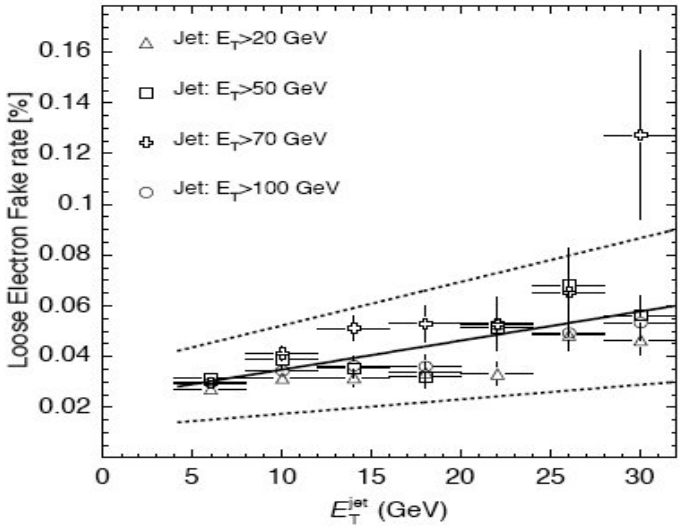
- Jets can pass electron ID cuts,
  - Mostly due to
    - early showering charged pions
    - Conversions: $\pi^0 \rightarrow \gamma\gamma \rightarrow ee+X$
    - Semileptonic b-decays
  - ♣ Difficult to model in MC
    - Hard fragmentation
    - Detailed simulation of calorimeter and tracking volume
- Measured in inclusive jet data at various E_T thresholds
  - Prompt electron content negligible:
    - N_{jet}~10 billion at 50 GeV!
  - **Fake rate per jet:**

	CDF	ATLAS
Loose cuts	$5x10^{-4}$	$5x10^{-3}$
Tight cuts	$1x10^{-4}$	$1x10^{-5}$

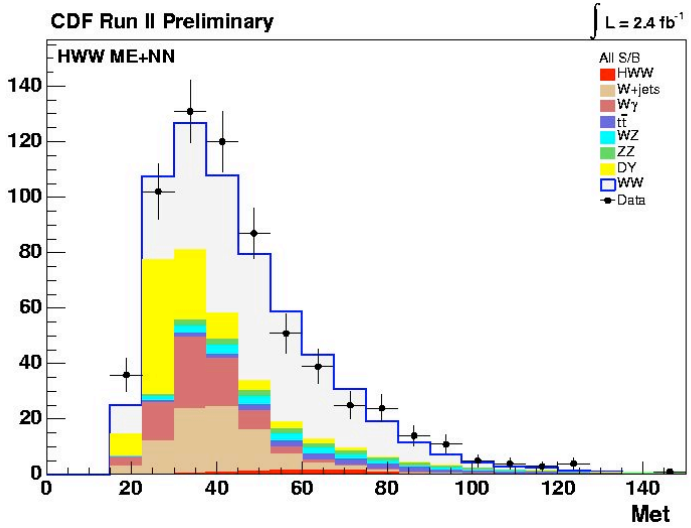
  - Typical uncertainties 50%

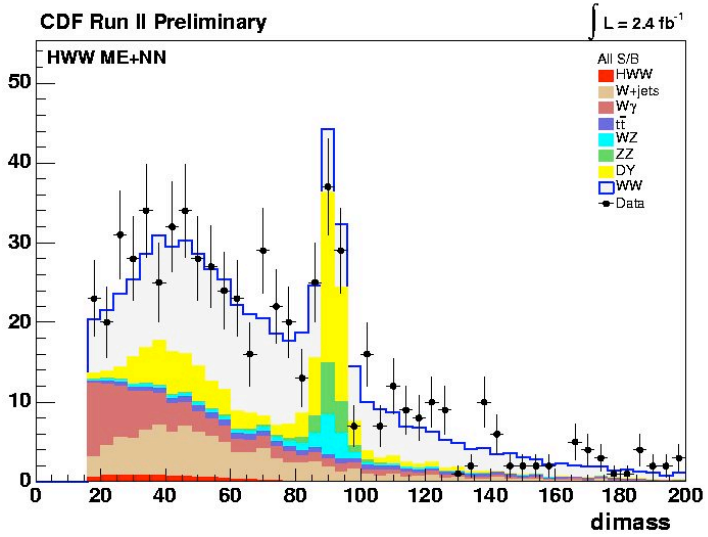


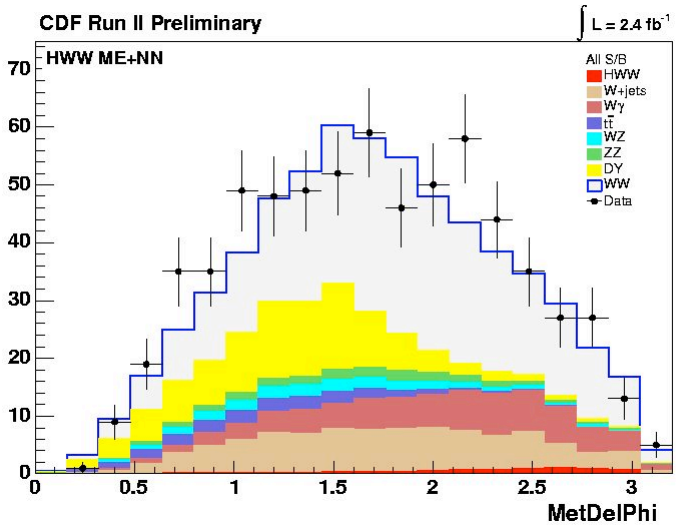
Fake Rate (%)

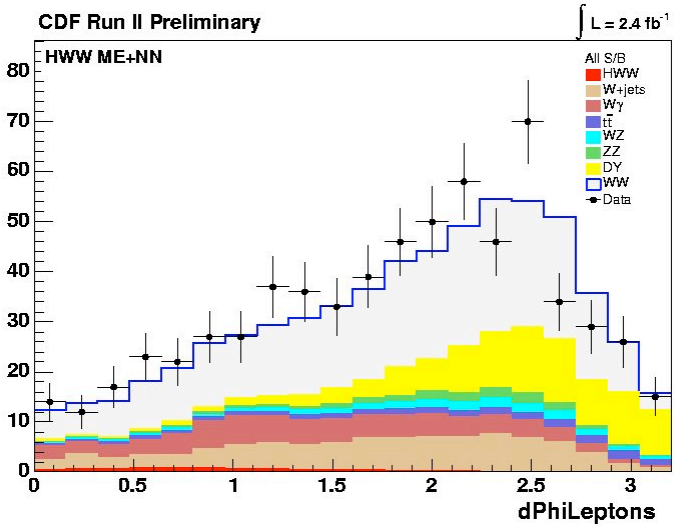


Plot Everything Under the Sun..



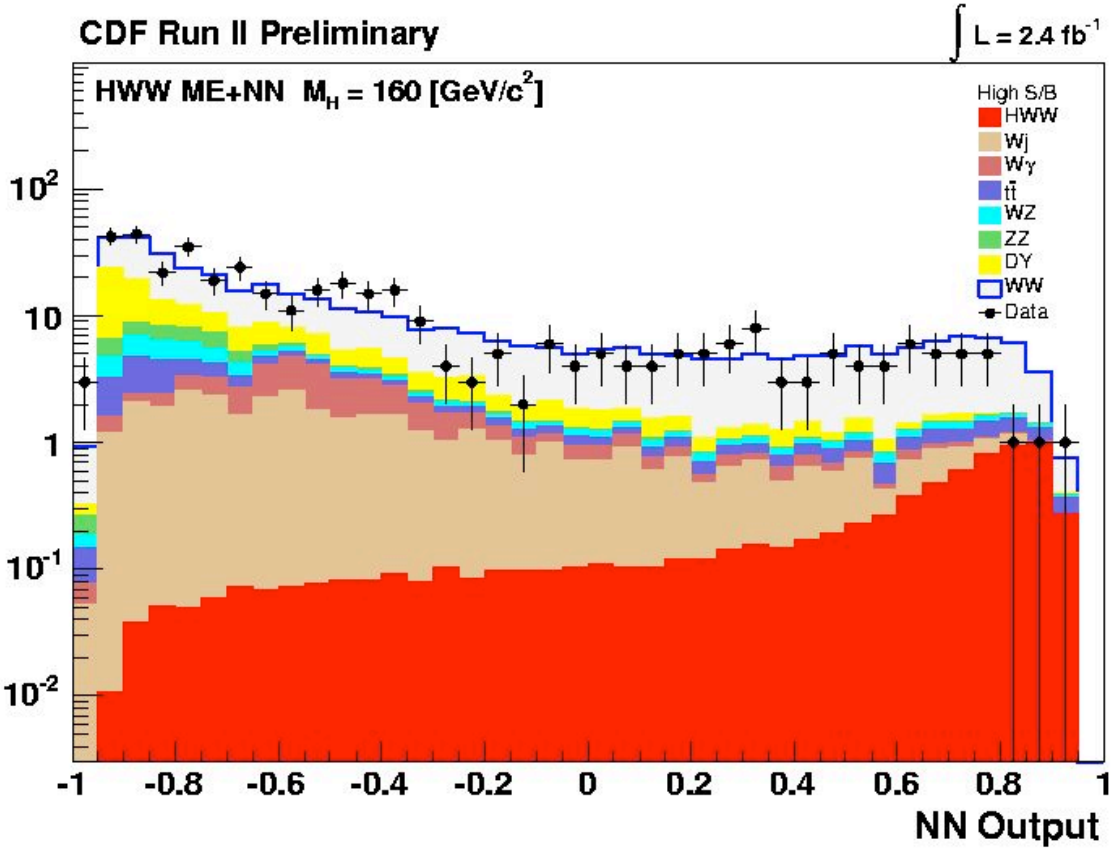






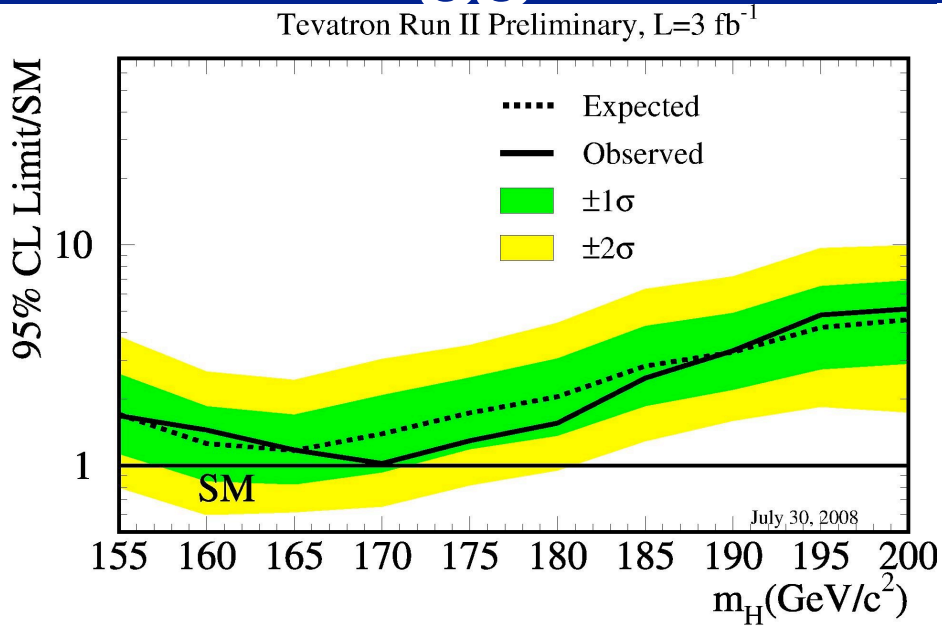
- **Validates the background prediction**
  - **Very often these plots "don't work" since there is some problem**
  - **Now plug all into sophisticated techniques!**

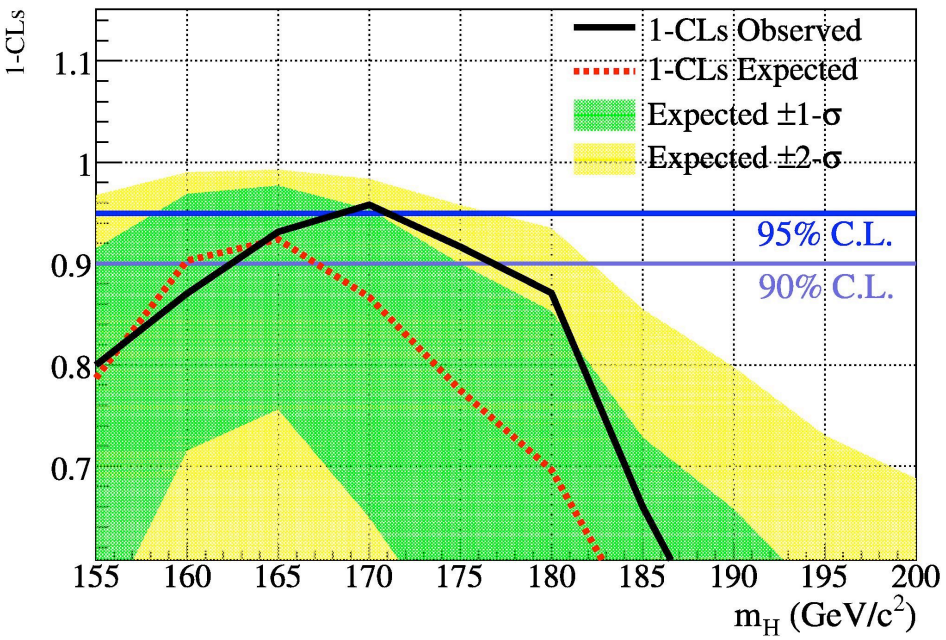
Result of Neural Network



- At high NN Output data fluctuated low
  - No sign of Higgs boson

Tevatron limits on the Higgs boson cross section

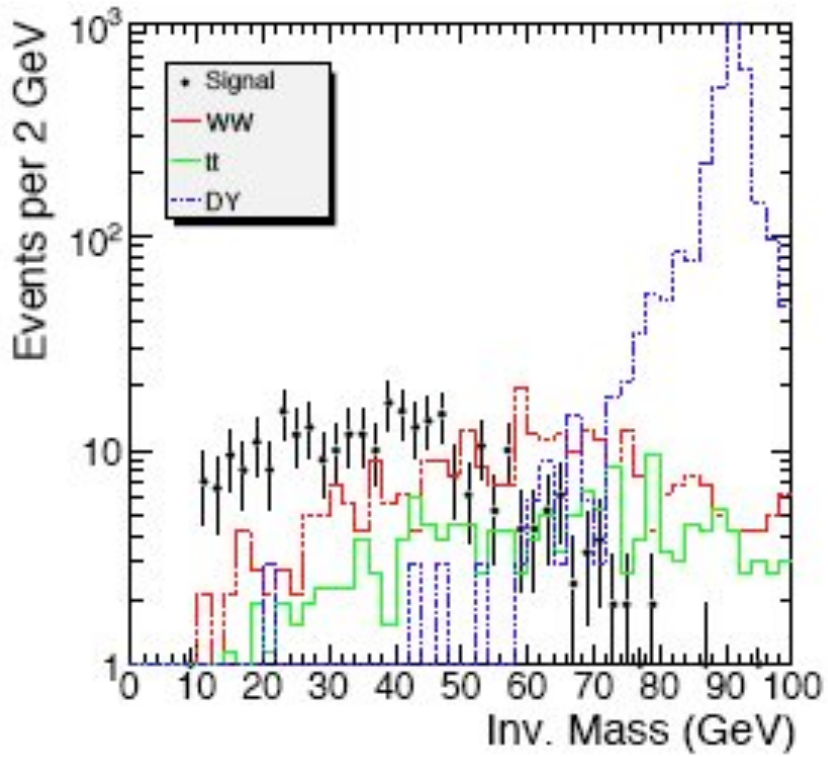


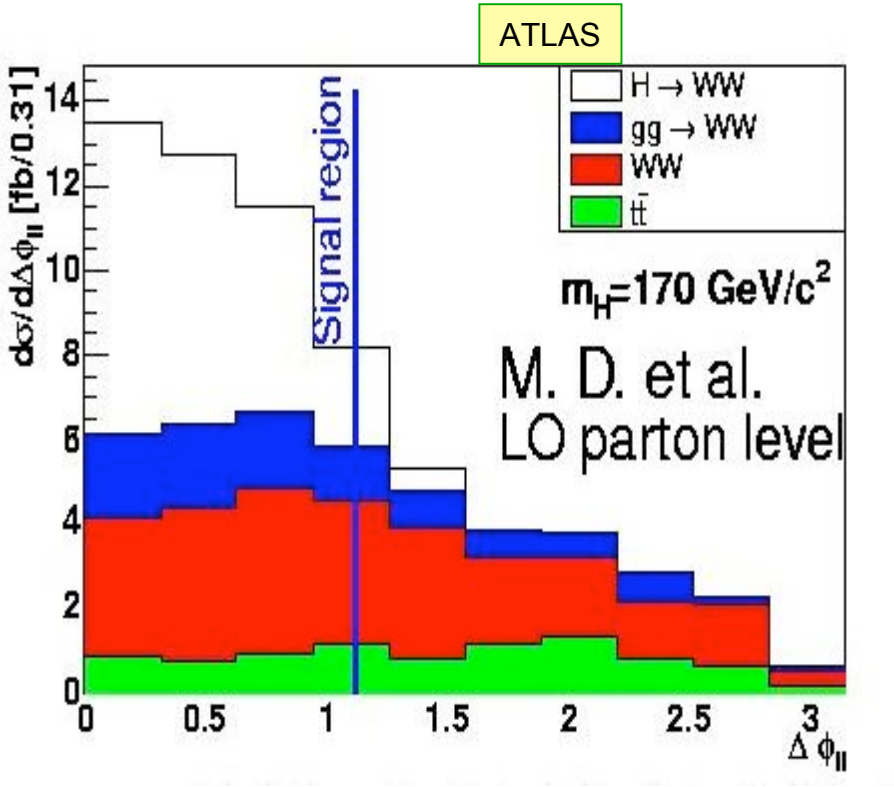


- Lack of observation
  - => an upper limit on the Higgs cross section
  - I.e. if the cross section was large we would have seen it!
- Results presented typically as ratio:
  - Experimental limit / theoretical cross section
  - If this hits 1 we exclude the Higgs boson at that mass!
- Brand new result: exclude M_H=170 GeV/c^2

Early Higgs Signals at LHC

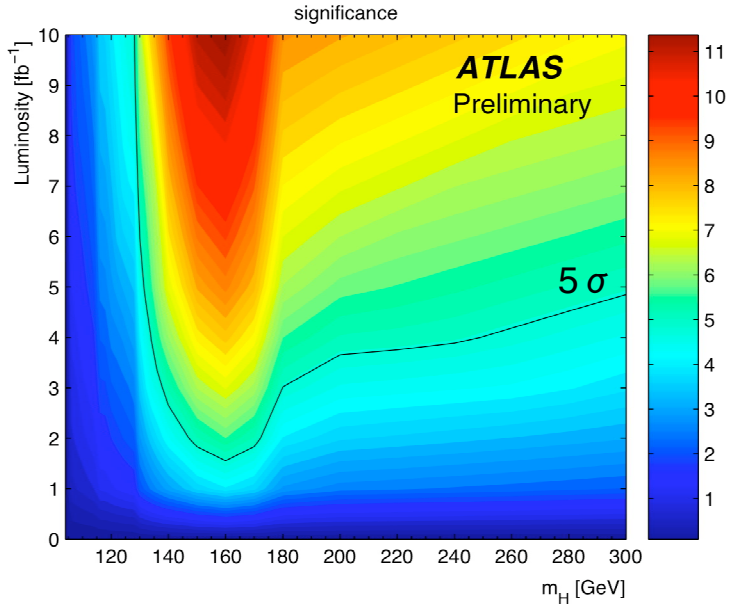
$H \rightarrow WW^*$ (m_H=170 GeV)

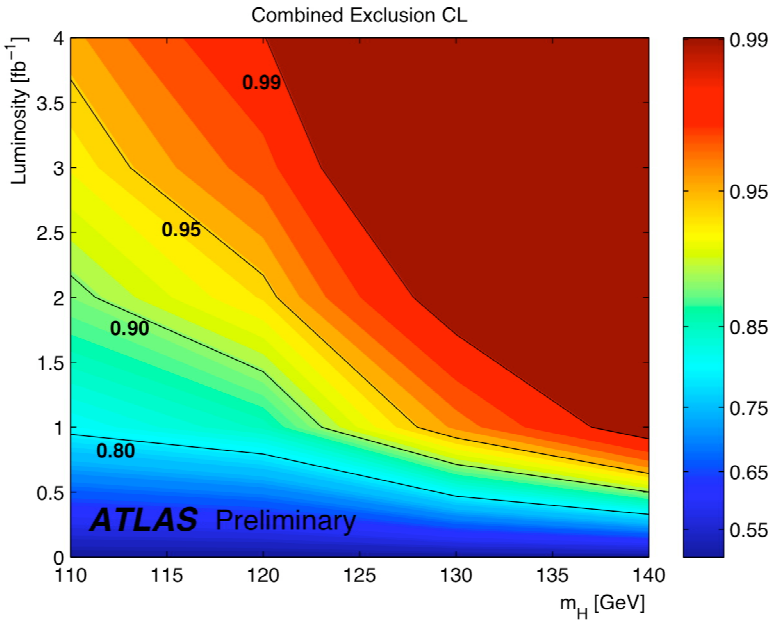


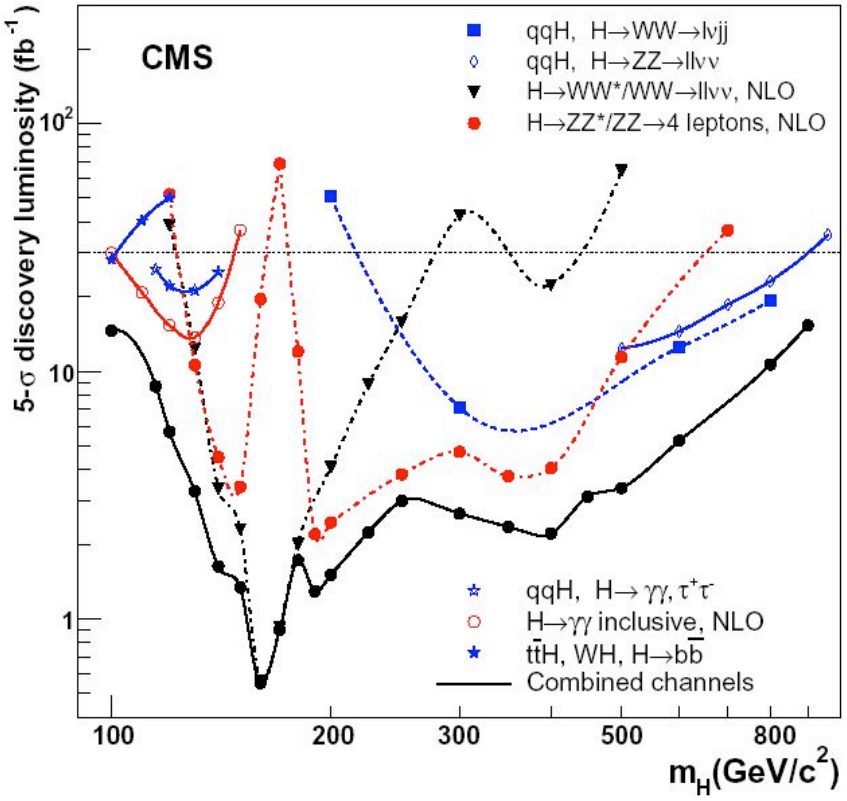


LHC has about 4 times better signal / background than Tevatron

LHC SM Higgs Discovery Potential







- 5σ discovery over full mass range with ~20 fb^{-1}
  - Most challenging at low mass
- 95% exclusion over full mass range with ~4 fb^{-1}

Concluding Remarks

Data are very precious

- Treat them with the highest respect
- Try to not jump to conclusions too fast
  - Data analysis is like detective work
- Try to use all you can to understand them
  - Redundancy of detector
    - tracker vs calorimeter etc.
  - Complementary physics processes
    - W's vs Z's etc.
  - Monte Carlo tools and theoretical calculations
- Use your brain and your judgment

This was my very personal view on the key issues of data analysis

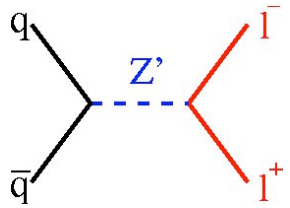
Thanks,

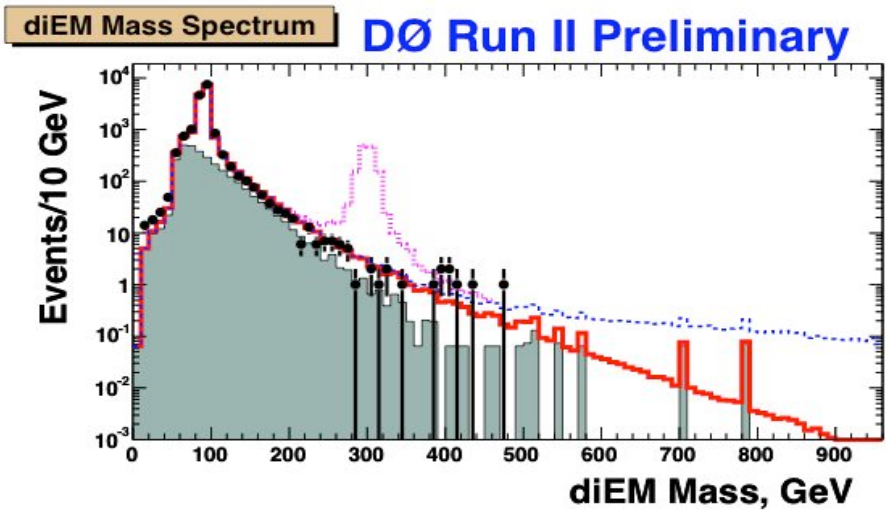
and lot's of fun and luck for your analyses

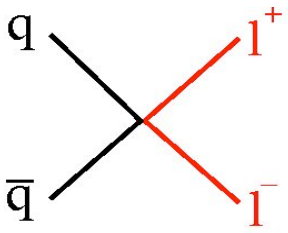
High Mass Resonances

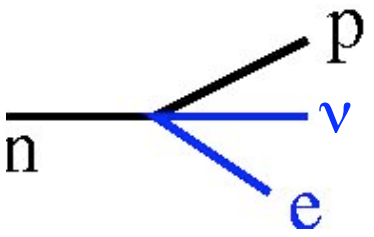
Resonances or Tails

- New resonant structure:
  - New gauge boson:
    - Z' →ee, μμ, ττ, tt
    - W' →eν, μν, τν, tb
  - Randall-Sundrum Graviton:
    - G→ee, μμ, ττ, γγ, WW, ZZ,…
- Tail:
  - Large extra dimensions (ADD model)
    - Many many many resonances close to each other:
    - "Kaluza-Klein-Tower": ee, μμ, ττ, γγ, WW, ZZ,…
  - Contact interaction
    - Effective 4-point vertex
      - E.g. via t-channel exchange of very heavy particle
    - Like Fermi's β-decay



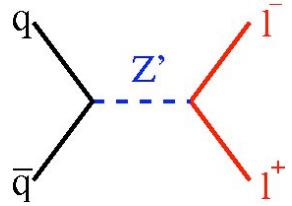


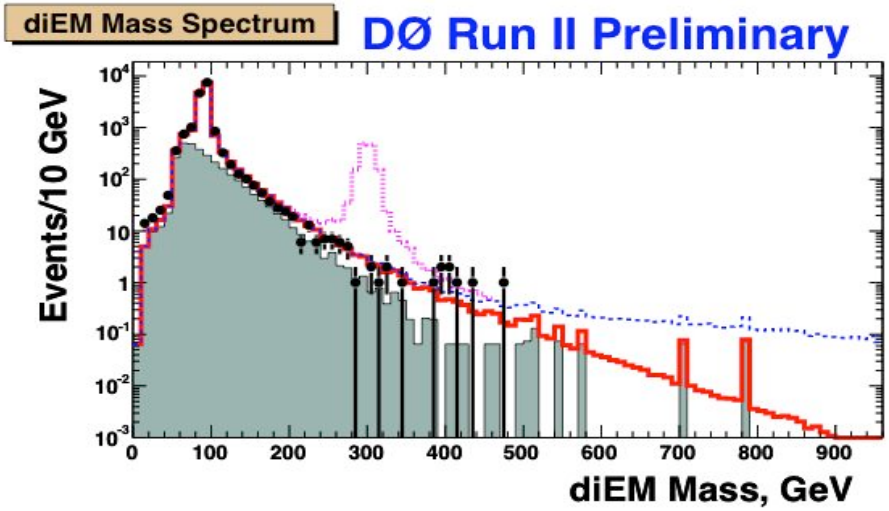


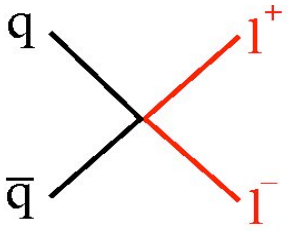


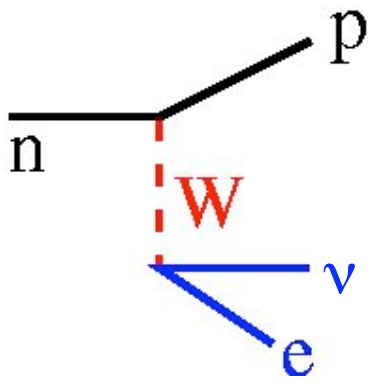
Resonances or Tails

- New resonant structure:
  - New gauge boson:
    - Z' →ee, μμ, ττ, tt
    - W' →eν, μν, τν, tb
  - Randall-Sundrum Graviton:
    - G→ee, μμ, ττ, γγ, WW, ZZ,…
- Tail:
  - Large extra dimensions (ADD model)
    - Many many many resonances close to each other:
    - "Kaluza-Klein-Tower": ee, μμ, ττ, γγ, WW, ZZ,…
  - Contact interaction
    - Effective 4-point vertex
      - E.g. via t-channel exchange of very heavy particle
    - Like Fermi's β-decay



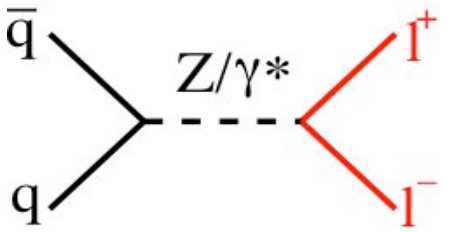


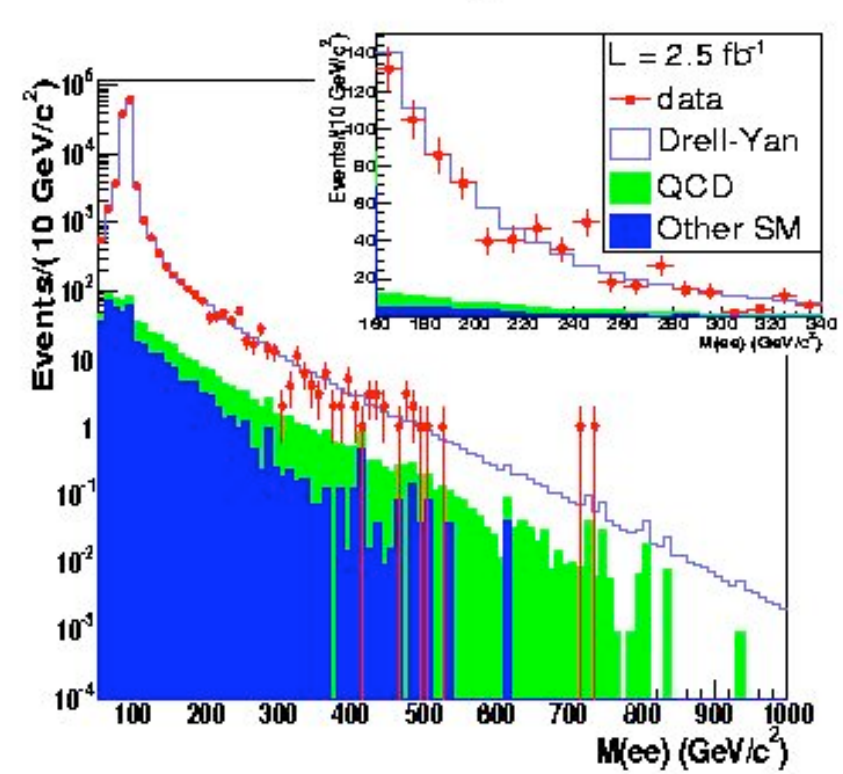




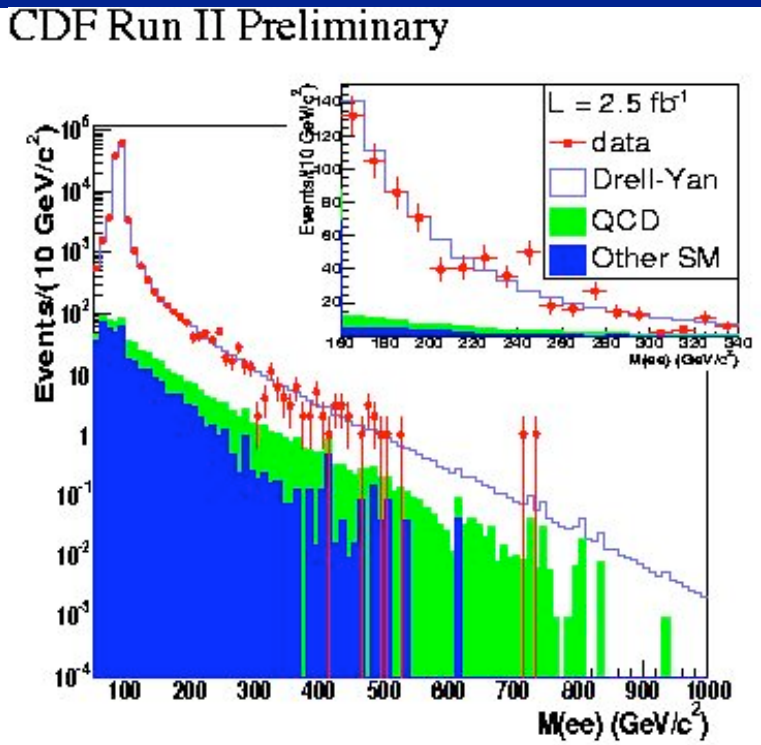
Dilepton Selection

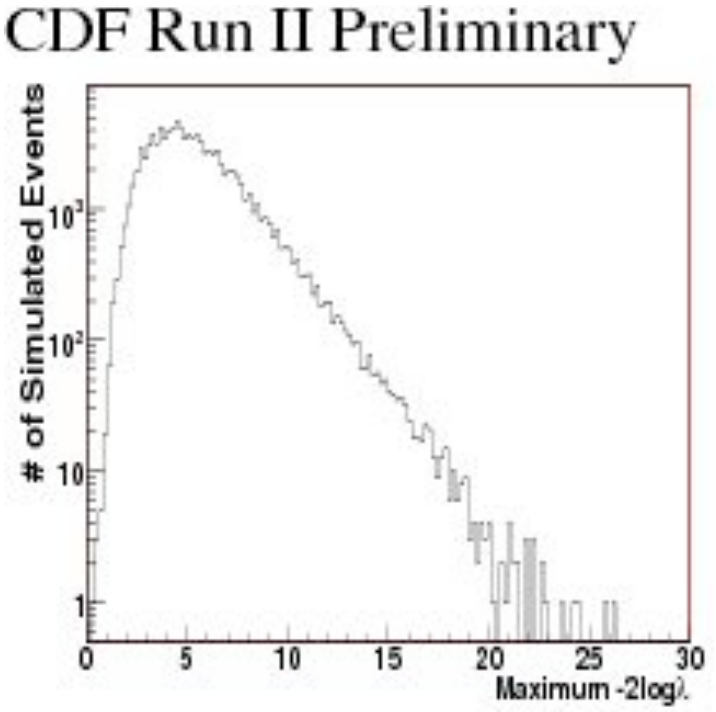
- Background to two high momentum leptons
  - irreducible background is Drell-Yan production
    - Well known/calculated by MC
    - NNLO calculation also available
  - Other backgrounds:
    - Jets faking leptons: reject by making optimal lepton ID cuts
    - WW, diphoton, etc. very small
- None of these backgrounds are expected to make peaks in mass distribution





Neutral Spin-1 Bosons: Z'

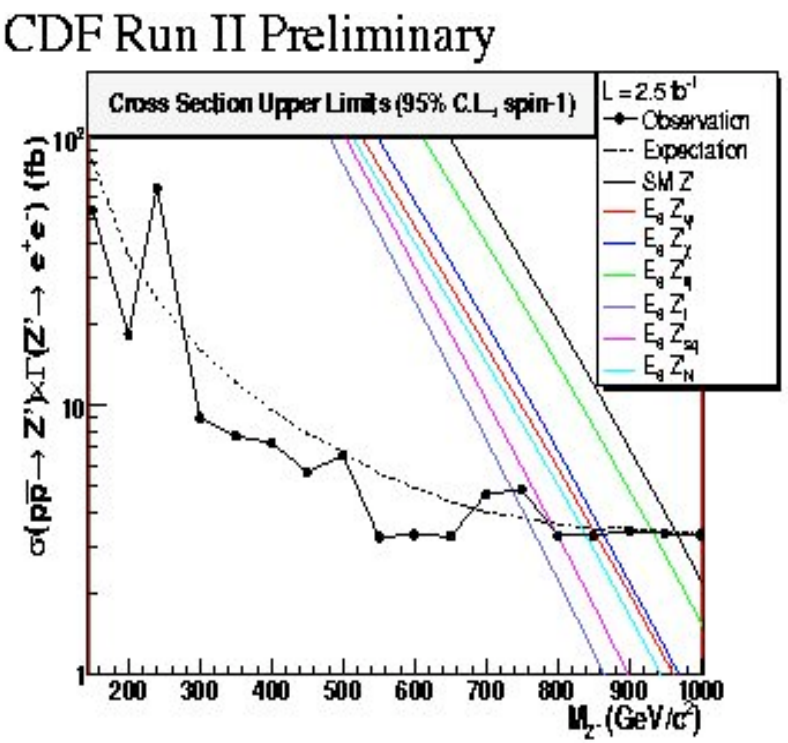




- 3.8σ excess at m(ee)=240 GeV/c^2
- But need to evaluate how likely this happens somewhere in the analysis!
- Technique:
  - Run Pseudo-experiments to see how often statistically on gets fluctuation of >3.8
- Result: 0.6% probability (2.5 σ)
  - Need to see how new data look like

Interpreting the Mass plots

- No evidence for any deviation from Standard Model => Set limits on new physics
  - Set limits on cross section x branching ratio
    - This is model independent, i.e. really what we measure
    - Any theorist can overlay their favorite curve
    - It remains valid independent of changes in theory
    - Always publish this!
    - Here: σ<4 fb at high mass
  - Can also set limits on Z' mass within certain models
    - This is model dependent
    - Nice though for comparing experiments, e.g. LEP vs Tevatron



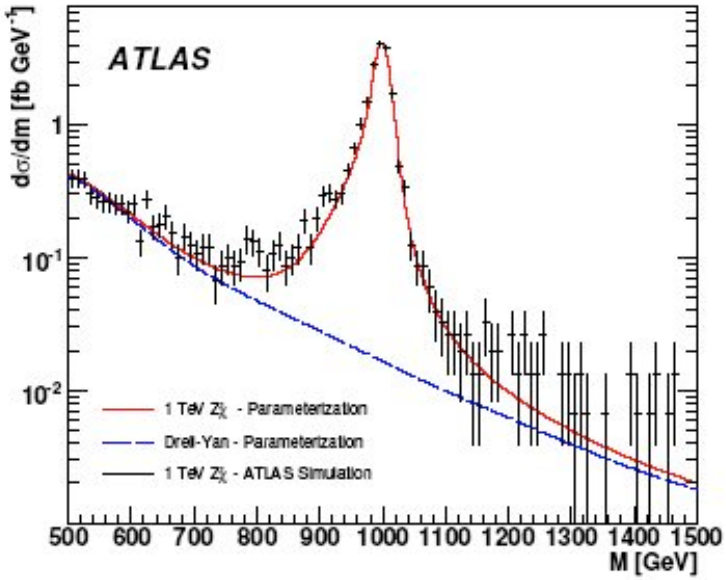
For SM couplings:

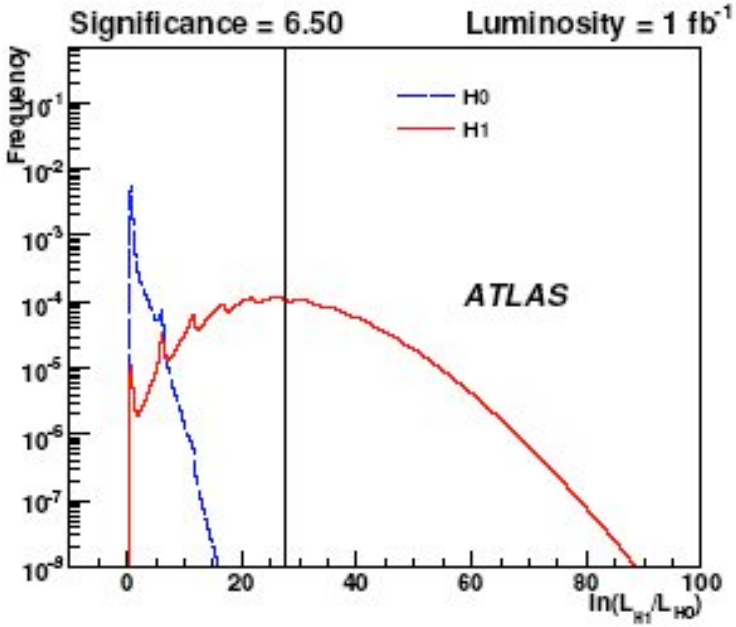
	Z′→ee
limit	M>966 GeV

Alternative Method

- Fit for bump in mass spectrum:
  - Parameterize distribution
  - Convolute with efficiency and detector resolution

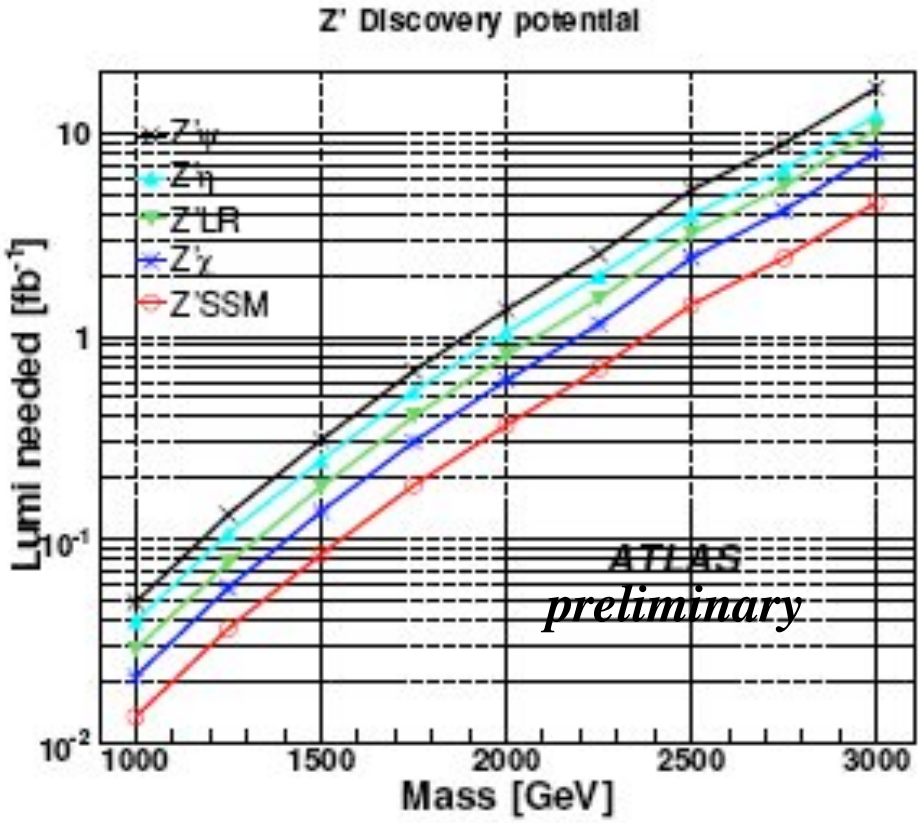
$$\left.\frac{d\sigma}{dm}\right|_{\text{Signal}}(m) = \frac{1}{m^2} \times G_{PDF}(m) + \mathscr{A}_{\text{peak}} \times \frac{\Gamma_{Z'}^2}{m_{Z'}^2} \frac{m^2}{(m^2 - m_{Z'}^2)^2 + m_{Z'}^2 \Gamma_{Z'}^2} \times G_{PDF}(m) + \mathscr{A}_{\text{interf}} \times \frac{\Gamma_{Z'}^2}{m_{Z'}^2} \frac{m^2 - m_{Z'}^2}{(m^2 - m_{Z'}^2)^2 + m_{Z'}^2 \Gamma_{Z'}^2} \times G_{PDF}(m)$$





Significance given by ratio of likelihood of signal-hypothesis vs likelihood of null-hypothesis

LHC Z´ Discovery Potential



- Probe >1 TeV already with 100 pb^{-1}
- Improves by ~1 TeV for every factor 10 in luminosity
- Systematic uncertainties relatively negligible in these analyses